TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

MATIAS IJÄS

LASKENNALLINEN ALGORITMI SOSIAALISESTI SYRJÄÄNTYNEIDEN LÖYTÄMISEKSI SOSIAALISISTA VERKOSTOISTA

Kandidaatintyö

Tarkastaja: Maria Törhönen
Tarkastaja ja aihe hyväksytty
27.8.2018

TIIVISTELMÄ

MATIAS IJÄS: Laskennallinen algoritmi sosiaalisesti syrjäytyneiden löytämiseksi sosiaalisista verkostoista
Tampereen teknillinen yliopisto
Kandidaatintyö, 42 sivua, 1 liitesivu
Joulukuu 2018
Tietotekniikan diplomi-insinöörin tutkinto-ohjelma
Pääaine: Ohjelmistotuotanto
Tarkastaja: Maria Törhönen

Avainsanat: Verkkomallinnus, Algoritmi, Sosiaaliset verkostot, Syrjäytyneet, Simulaatiomalli, Sosiaalinen syrjäytyminen

Syrjäytyminen yhteiskunnasta ja sosiaalisesta kanssakäymisestä on puhuttanut mediassa ja valtion ylimmässä johdossa. Sitä on pidetty ongelmallisena ja suurena huolenaiheena koskien pääosin nuoria. Syrjäytyminen ei rajoitu pelkästään nuorisoon, vaan sitä esiintyy jo pienillä lapsilla yksinäisyyden ja ryhmään kuulumattomuuden muodoissa. Sitä esiintyy myös aikuisilla esimerkiksi työttömyyden tai masentuneisuuden myötä. Sosiaalisista suhteista syrjäytymistä kuvataan vieraantumisella.

Tässä työssä esittelen uuden algoritmin syrjäytyneiden löytämiseksi sosiaalisista verkostoista käyttäen verkkomallinnuksen periaatteita. Algoritmi käyttää tietoa ryhmittymien muodostumisesta sekä yksittäisten solmujen lasketusta vaikutusvallasta. Algoritmin teoria pohjautuu paikallisten maksimien löytämiseen, verkkotopologian yksityiskohtaiseen läpikäymiseen sekä matemaattisiin malleihin, joiden avulla vaikutusvaltaa mitataan. Vaikutusta mitataan laskemalla solmujen ja linkkien painoarvoja, laskemalla levinneisyyttä yhdestä solmusta kaikkiin ja kaikista yhteen. Algoritmin eri komponentteja voidaan käyttää moneen tarkoitukseen, joista syrjäytyneiden etsiminen on yksi sovelluskohde. Algoritmin laskennallinen tehokkuus ja sen luotettavuus realistisissa tapauksissa osoitetaan.

Algoritmin tuottamia tuloksia on tarkasteltu ja niistä on huomattu loogisia yhtäläisyyksiä verkon topologiaan verrattuna. Algoritmi sopii tilanteisiin, joissa verkon topologia voidaan mallintaa reaalimaailmasta. Luokkahuoneen topologia voidaan mallintaa esimerkiksi pyytämällä lapsia nimeämään kavereitaan kyseiseltä luokalta. Algoritmin tarkoituksena on tunnistaa syrjäytyviä yksilöitä ajoissa ja täten ehkäistä syrjäytymisestä ja yksinäisyydestä aiheutuvia haittavaikutuksia tunnetuissa sosiaalisissa ympäristöissä. Syrjäytymisen ehkäisemiseksi yhteiskunnallisessa mielessä algoritmi auttaa havaitsemaan yksilöt, joilla ei ole laajaa verkostoa. Työ- tai opiskelupaikka olisi mahdollista saada helpommin kontaktien avulla.

ABSTRACT

MATIAS IJÄS: Computational Algorithm for Finding Outcasts from Social Networks
Tampere University of Technology
Bachelor of Science Thesis, 42 pages, 1 Appendix page
December 2018
Master's Degree Programme in Information Technology
Major: Software Engineering
Examiner: Maria Törhönen

Keywords: Network analysis, Algorithm, Social networks, Outcasts, Simulation model, Social exclusion

Social isolation from the society and from social interactions has been a topic in the media, including the government's highest leaders. It has been considered problematic and major concern for mostly young people. However, the exclusion is not limited only towards young people, but it is already present in lives of small children in forms of loneliness and non-belonging. It is also present in adult lives, for example, due to unemployment or depression. The social isolation is described by the term alienation.

In this research, I present a new algorithm for finding excluded people in social networks using the principles of network modeling. The algorithm uses information about the formation of communities and the computed influence of individual nodes. The theory of the algorithm is based on locating the local maxima, going through the detailed topology of the network and mathematical models measuring influence. Influence is measured by computing the weights of nodes and links, by computing spreading probabilities from a single node to everywhere and from all nodes to the selected node. Different modules of the algorithm can be used for many purposes, from which searching of outcasts is one application. The efficiency of the algorithm and its reliability in realistic cases will be demonstrated.

The results of the algorithm have been studied and logical similarities over the network topology have been found. The algorithm is suitable for situations where network topology can be modeled from the real world. Classroom topology can be modeled for example by asking children to name their friends in the class. The purpose of the algorithm is to find the isolated persons to prevent the negative effects of exclusion and loneliness in well-known social environments. To prevent social exclusion from the view of society, the algorithm helps to detect individuals without a large network. It would be easier to get a job or study place through contacts.

ALKUSANAT

Tämä tekniikan kandidaatintyö on tehty Puolustusvoimien tutkimuslaitokselle ja se on hyväksytetty Tampereen teknillisen yliopiston Tietotekniikan laitoksella.

Kiitän suuressa roolissa algoritmin kehittämisessä mukana ollutta Janne Levijokea. Haluan kiittää Puolustusvoimissa ohjaajanani toiminutta INS Jarkko Karista sekä verkkomallinnuksessa opastanutta erikoistutkija FL Vesa Kuikkaa. Lisäksi haluan kiittää Puolustusvoimien tutkimuslaitosta mielenkiintoisesta ja mukavasta työympäristöstä. Kiitän myös Tampereen teknillisellä yliopistolla tarkastajana toiminutta Maria Törhöstä.

Tampereella, 09.12.2018

Matias Ijäs

SISÄLLYSLUETTELO

KUVALUETTELO

TAULUKKOLUETTELO

OHJELMALUETTELO

LYHENTEET JA MERKINNÄT

GB	gigatavu on mitta muistimäärälle, engl. Gigabyte
O()	O-notaatio, laskennan tehokkuuden merkintätapa
SAW	itseään välttelevä polku, engl. Self-avoiding (random) walk
SNA	Sosiaalisen verkon analysointi, engl. Social Network Analysis
SNAP	Stanford Network Analysis Project, avoimen datan lähde
$C(i)$	keskeisyyden mitta (yleinen merkintä), engl. Centrality measure
$C_b(i)$	välittäjyyden mitta, engl. Betweenness centrality
$C_b'(i)$	normalisoitu välittäjyyden mitta, lukuarvot nollasta yhteen
$C_c(i)$	keskeisyyden mitta, engl. Closeness centrality
$Coh(G)$	koheesio ryhmälle G
$Coh_i(G_1, G_2)$	koheesioindeksi, summattu koheesio ryhmille G_1 ja G_2. Kuvaa kahtiajaon kokonaiskoheesiota
$\Delta Coh_i(G_1, G_2)$	muutos koheesioindeksiin
G	ryhmä
$Gmax$	ryhmä, jonka todennäköisyys on suurin, ryhmän koko $N > 1$
$Gr_i(n)$	Ryhmään kuulumisen todennäköisyys solmulle n
ΔGr_p	mitta eroavaisuudelle ryhmän ja henkilön aatteiden välillä
$g_{i,j}$	geodeettinen etäisyys (lyhin polun pituus solmujen i ja j välillä)
i	indeksi (kuvaa solmun indeksiä tai solmua)
j	indeksi (kuvaa solmun indeksiä tai solmua)
L	polun pituus
L^{Gr}	ryhmän käyttäytymis- tai aatemallia kuvaava lukuarvo
L_{max}	maksimi polun pituus (käyttäjän valitsema)
L^P	henkilön käyttäytymis- tai aatemallia kuvaava lukuarvo
N	lukumäärä (solmujen lukumäärä verkossa)
N_{L_1}	lista solmuista, jotka kuuluvat polkuun L_1
$Norm(L)$	normaalijakauman arvo polun pituudelle L
n	solmu (solmun indeksi tai sisältö)
$O_i(n)$	syrjäytymisen todennäköisyys solmulle n, engl. Outcast index
$P_{s,t}$	leviämistodennäköisyys lähtösolmusta s kohdesolmuun t
$Po(L)$	Poisson-jakauman arvo polun pituudelle L
p_{Gmax}	todennäköisimmän ryhmän todennäköisyys
p_L	polun L todennäköisyys (polun pituus L)
p_L^*	polun L painottamaton todennäköisyys (polun pituus L)
p_l	yhteisen polun l todennäköisyys
p_n	solmun yksinolemisen todennäköisyys ryhmässä
S_{Gmax}	suurimman todennäköisyyden omaavan ryhmän koko
S_{Gr}	ryhmän sympatia erilaisuutta kohtaan
S_n	yksittäissolmun ryhmän koko, mikäli solmu voi esiintyä yksin, $S_n = 1$
s	indeksi (kuvaa lähtösolmua, engl. source)
T	aika, kuvataan mm. Poisson jakaumassa aikariippuvuutta
$Tasa(L)$	tasajakauman arvo polun pituudelle L
t	indeksi (kuvaa kohdesolmua, engl. target)
Y_{L_1}	lista yhteyksistä (linkeistä), jotka kuuluvat polkuun L_1
y	yhteys, linkki, engl. Connection, edge

$w\_Tasa(L)$	painotetun tasajakauman arvo polun pituudelle L
w_n	painoarvo solmulle n, engl. node weight
w_y	painoarvo linkille y, engl. edge weight
λ	parametri Poisson jakaumassa, joka säätelee kumulatiivisen todennäköisyyden raja-arvon 1 saavuttamista ajanhetkellä T
μ	parametri, kuvaa normaalijakauman keskipistettä
σ_{st}	lyhimpien polkujen summa solmusta s solmuun t
$\sigma_{st}(i)$	lyhimpien polkujen summa solmusta s solmuun t solmun i kautta
σ^2	parametri, kuvaa normaalijakauman varianssia

SANASTO JA MÄÄRITELMÄT

Attribuuttidata	Solmukohtainen data, sisältää esimerkiksi sukupuolen ja iän
Big data	Tietomassa, jota on paljon, sisältää vaihtelua ja vaatii paljon tehoa laskennassa algoritmin ja/tai laitteiston osalta
Geodeettinen etäisyys	Lyhin polun pituus kahden solmun välillä
Eristäytyminen	Sosiaalisesta ympäristöstä omalla valinnalla vieraantunut
Keskeisyys	Mitta vaikuttavuudelle verkostossa
Koheesio	Ryhmän tiiviyttä kuvaava suure
Limittäinen yhteisö	Yhteisö, jolla on samoja ja eri solmuja toisen yhteisön kanssa
Linkki	Yhdistää kahta solmua, sisältää painoarvon väliltä 0 ja 1
Mallintaminen	Perustuu malliin, jolla pyritään kuvaamaan tosielämää
Naapurisolmu	Solmu, joka on yhden linkin päässä valitusta solmusta
Poisson-jakauma	Käytetty matemaattinen kaava ajankulun määrittämiseksi
Päällekkäinen yhteisö	Yhteisö, jossa on vain samoja solmuja toisen yhteisön kanssa eli on toisen yhteisön osajoukko.
Rihmasto	Verkko, jossa on pinnan alla olevia (ei-tiedettyjä) yhteyksiä
Ryhmittymä	Suuremman yhteisön sisällä oleva tiivis pienryhmä
Simulaatio	Matemaattinen näkemys tilanteesta, jota ajetaan tietokoneella lukuisia kertoja, jotta erilaisia skenaarioita voidaan arvioida
Solmun yhteysaste	Solmuun tulevien ja solmusta lähtevien linkkien summa
Solmu	Yhteyspiste, joka linkkien avulla muodostaa verkoston
Sosiometria	Laskennallinen metodi sosiaalisten suhteiden mittaamiseen
Syrjäytynyt	Yhteiskunnallisesta tai sosiaalisesta ympäristöstään erillään oleva henkilö, jonka suhteet muihin ovat heikkoja
Topologia	Verkoston solmujen ja linkkien luoma avaruus keskeisyyden perusteella, jossa joka solmulle on abstrakti paikka.
Verkkomallinnus	Matemaattisia malleja hyödyntävä menetelmä verkostojen analysoimiseen
Verkon rakenne	Muodostuu solmujen yhteyksistä ja niiden painoarvoista
Verkosto	Verkko, jonka rakenne on tiedossa, eroaa hieman rihmastosta
Vieraantunut	Sosiaalisesta ympäristöstä erillään oleva henkilö
Välittäjyys	Mitta tiedon levittämiseen verkostossa
Yhteisöjen etsiminen	Verkkomallinnuksen alalaji, jossa verkostosta etsitään ryhmiä
Yhteysdata	Linkkikohtainen data, sisältää vaikuttavat solmut, painoarvon

1. JOHDANTO

Verkkomallinnus (engl. Network Science) on vuosien saatossa kehittynyt siihen pisteeseen, että sitä voidaan hyödyntää tutkimuksessa. Sen tavoitteena on antaa nopeita ratkaisuja ja selvittää pinnan alla olevia tekijöitä erilaisten verkostojen taustalla. Monet tosielämän rakenteet voidaan mallintaa verkostoiksi, jossa yhteyspisteinä eli solmuina (engl. Node) voivat toimia esimerkiksi ihmiset, signaalinvälittäjät tai sähköverkon osat ja yhteyksinä toimivat linkit (engl. Edge), kuten esimerkiksi vaikutus, signaalit tai sähkö [78].

Erilaisia mittalukuja keskeisyyden (engl. Closeness centrality) ja välittäjyyden (engl. Betweenness centrality) mittaamiseksi on esitetty kirjallisuudessa [28][59]. Myös useita tapoja etsiä ryhmittymiä (engl. Subgroups, Clusters) on esitetty kirjallisuudessa [3][14][26][38]. Näihin algoritmeihin on kuitenkin tehty muutoksia, jotta vaikuttavuus (engl. Influence) saadaan laskettua jokaisesta solmusta jokaiseen [39] ja ryhmittymien etsiminen skaalautuisi O(N^2) luokkaan [46].

Yleiseksi trendiksi matemaattisia malleja tehdessä on noussut big data. Algoritmien hyvä skaalautuvuus (engl. Scalability), etenkin O-notaatiolla mitattuna on noussut suureen arvoon. Tämän työn vaikuttamisen malli pohjautuu kompleksisiin mallinnusmenetelmiin (engl. Complex network) [44], joissa laskettiin kaikki mahdolliset polut lähtösolmun ja maalisolmun välillä. Aiemmassa algoritmissa [46] polkujen lukumäärän kasvu tapahtui kuitenkin eksponentiaalisesti koko verkoston kasvuun suhteutettuna, jolloin vaikuttamisen algoritmi skaalattiin lineaarisiin lukemiin [39].

Syrjäytyneisyys on noussut etenkin viime vuosina suureksi huolenaiheeksi päättäjien keskuudessa. Syrjäytyneisyyden määritelmään kuuluu suhteiden heikkeneminen ihmissuhteissa ja yhteiskuntaa kohtaan, muun muassa työttömyyden takia [19][35]. Aihetta tutkimalla luotiin karkea malli, jossa syrjäytymisen riskiä analysoitiin. Tässä algoritmissa yhdistetään vaikuttavuuden erilaiset muodot ja potentiaaliset ryhmittymät, joiden pohjalta lasketaan todennäköisyydet henkilöiden sosiaaliselle syrjäytyneisyydelle. Tämän jälkeen tarkastellaan potentiaalisia syrjäytyneitä yksittäin ja arvioidaan riski syrjäytyneisyydelle muun muassa ryhmään kuuluvuuden indeksillä.

Euroopan unionin alueella on yli 100 miljoonaa ihmistä, jotka ovat vaarassa kuulua köyhyyden tai syrjäytyneisyyden piiriin. Suurimmassa riskissä ovat Romania ja Bulgaria yli 40 %:n todennäköisyydellä. Suomella ja Ruotsilla todennäköisyys on alle 20 % [20]. Todellinen ja havaittu sosiaalinen eristäytyminen on yhteydessä ennenaikaisen kuoleman

riskiin. Objektiivisen ja subjektiivisen sosiaalisen eristäytymisen vaikutus on verrattavissa vakiintuneeseen riskiin kuolla suuremmalla todennäköisyydellä ennenaikaisesti [37].

Tämän kandidaatintyön alussa, luvussa 2, esitellään tarkempaa kirjallista taustaa verkkomallinnuksesta, matemaattisesta mallista, algoritmissa käytetyistä kaavoista ja ryhmittymien muodostumisesta sekä syrjäytyneisyydestä. Luvussa 3 käydään teorian ja matematiikan osalta läpi pohjana käytetty malli vaikuttamisesta, sekä malliin tehdyistä muutoksista. Luvussa on myös esitelty tehty teoriapohjainen toteutus ryhmien muodostumisesta ja syrjäytyneisyyden hakemiselle näiden tietojen avulla. Vertailua tehdään vastaaviin kirjallisuudessa jo dokumentoituihin toteutuksiin.

Luvussa 4 syvennytään algoritmien toimintaan ohjelman tasolla. Vaikuttamismallia kuvaavia, ryhmittymien muodostuksessa käytettyjä sekä syrjäytymiseen sovellettuja algoritmeja tarkastellaan skaalautuvuuden, tehokkuuden ja tarkkuuden mukaan. Luku 5.1 käsittelee pieniä verkostoja. Alaluvuissa 5.1.1 ja 5.1.2 käytetään reaalimaailman verkostoja hollantilaisopiskelijoista ja karateklubista, joista esittellään datan perusteella löytyviä syrjäytyneitä ja mahdollisesti syrjäytyneitä henkilöitä. Näihin ohjelma antaa arvion riskialueella oleville.

Luku 5.2 käsittelee trendinä olevaa big dataa, jossa kerrotaan yleisesti sen ongelmista ja hyödyistä. Tämän lisäksi ohjelman eri osien toimintaa tarkastellaan big datan näkökulmista esimerkkiverkon avulla. Luku 6 kokoaa yhteen johtopäätökset tuloksista ja kertoo parannusehdotuksista sovellukselle sekä luetteloi tässä toteutuksessa käytetyiden algoritmien ja mallien muita sovelluskohteita.

2. TAUSTA

Ennen 1900-lukua Emile Durkheim ja Ferdinand Tönnies sivusivat sosiaalisten verkostojen ideaa teorioiden ja tutkimuksen keinoin [18][79]. Erilaisia verkkoja ja verkostoja on laajemmin tutkittu sosiaalitieteissä 1930-luvulta lähtien. Ihmisten välisten linkkien muodostama kokonaisuus eli rakenne (engl. Structure) muodostui tärkeäksi yhteiskunnalle [58]. 1930-luvun jälkeen on ollut käytössä kaksi selkeää käsitettä: Radcliffe-Brownin sosiaalinen rakenne (engl. Social structure) ja Morenon sosiometria (engl. Sociometry) [56][67]. Käsitteitä on toki tullut lisää 1960-luvulla ja sen jälkeen.

Verkkomallinnuksen ja verkostojen syvällisemmän ymmärryksen saavuttamiseksi alaluvussa 2.1 tarkastellaan verkkomallinnusta nykyisellä termistöllä. Siinä käydään läpi, mistä verkkomallinnus koostuu, minkälaista dataa sillä analysoidaan ja mihin se soveltuu. Alaluvuissa 2.2 ja 2.3 tarkastellaan kirjallisuutta vaikuttamismalleista sekä ryhmittymien etsimisestä. Alaluvussa 2.4 käsitellään psykologiselta kannalta syrjäytymistä sekä siihen johtavia tekijöitä ja sen historiaa ja esitetään mittari, joka kuvaa yksilön kuuluvuutta ryhmään.

2.1 Verkkomallinnus

Graafiteoria eli verkkoteoria (engl. Graph theory) on matemaattisia malleja hyödyntävä osa-alue, joka tutkii alkioiden välisiä suhteita. Graafiteoria sai alkunsa Eulerin ratkaisusta Königsbergin silloista 1736 [22]. Graafiteoriaa voidaan soveltaa eri tieteenaloilla [31], kuten sähkötekniikassa [78], sosiologiassa [33] ja biologiassa [52]. Usein verkot koostuvat solmuista ja linkeistä eli solmuja yhdistävistä kaarista tai yhteyksistä. Verkko voi olla sisältämättä solmuja tai linkkejä, mutta tällöin puhutaan yksinkertaisista tapauksista [32], joita kaikki eivät pidä verkkoina [58][80].

Solmuja ja linkkejä voi olla monentyyppisiä, esimerkiksi yksi solmu kuvaa miestä, toinen naista ja kolmas muun sukupuolen edustajaa. Linkki voi olla erilainen johtuen siitä, mitä se kuvaa, esimerkiksi ihmisten välisissä yhteyksissä yhdenlainen linkki voi tarkoittaa sosiaalisen median ystävää, toinen pelikaveria ja kolmas sydänystävää. Linkit voivat olla yksi- tai kaksisuuntaisia (engl. Directed, undirected) perustuen yhteyden laatuun. Solmut ja linkit voivat sisältää erilaisia painotuksia, ja niillä voi olla erilaisia ominaisuuksia [41]. Verkko voi olla täyden tiedon verkko eli verkosto, jossa kaikki yhteydet tiedetään. Verkostoja ovat esimerkiksi sosiaalisen median palvelut ja muut verkot, joista yhteysdata on saatavissa. Rihmasto on verkko, jossa osa yhteyksistä puuttuu ja sitä on täten vaikeampi arvioida. Rihmasto on esimerkiksi salaisten järjestöjen verkot, sillä tietoa niiden jäsenistä ja yhteyksistä ei ole saatavissa.

Sosiaalisten verkostojen analysoiminen (engl. Social network analysis, SNA) on kehittynyt nopeasti lisääntyneen datan määrän ja sosiaalisen median kasvun myötä [34][70][71]. Facebook, Twitter ja vastaavat verkkosivut kehottavat käyttäjiään muodostamaan verkostoaan kavereiden ja seuraajien avulla [12]. Sosiaaliset verkostot käsitteenä sisältävät digitaaliset verkostot, reaalimaailman ihmissuhteet [43], poliittiset yhteydet ja järjestöt, taloudelliset liiketoimet ja maantieteelliset ryhmät asuinpaikan tai valtion mukaan [72]. Analysoimisessa käytettävä data jaetaan periaatteessa kahteen osaan: attribuuttidataan (engl. Attribute data) ja yhteysdataan (engl. Relational data) [72].

Attribuuttidata koostuu asenteista, mielipiteistä ja käyttäytymisestä, jotka kuvastavat yksilön kannalta tutkittavaa asiaa. Näiden soveltaminen malliin vaatii datan muuttamista lukuarvoihin. Yhteysdata koostuu solmuista, yhteyksistä ja niiden vahvuuksista. Verkostojen analysoimisessa yhteyksiä käsitellään eri yksilöiden välisten linkkien avulla. Vaikka verkkoanalyysiä pidetään kvantitatiivisena eli numeroilla esitettävänä tutkimusmuotona laskutoimitustensa takia, se koostuu silti kvalitatiivisista eli laadullisista mittatuloksista verkon rakenteen ja kehyksen kuvaamiseksi. [72]

2.2 Vaikuttamismallien teoriat

Vaikuttamismallit perustuvat matemaattisiin kaavoihin, joissa pyritään muodostamaan malli siten, että keskeisiä solmuja (engl. Central nodes) ovat ne, joilla on paljon yhteyksiä. Yhteydet ovat myös vahvoja, ja ne kohdistuvat muihin vaikutusvaltaisiin solmuihin. Mitä kauempana solmu on keskeisistä solmuista ja mitä vähemmän sillä on yhteyksiä muihin solmuihin, sitä vähemmän merkittävä solmu on [51]. Vaikuttamismalleissa käydään tarkemmin läpi keskeisyyttä (engl. Closeness centrality), välittäjyyttä (engl. Betweenness centrality) ja solmujen yhteysastetta (engl. Node degree).

Tärkeimpien solmujen löytäminen verkon rakenteesta on ollut merkittävänä tutkimuskysymyksenä kompleksisten verkkojen (engl. Complex networks) mallinnuksessa. Tutkimuskohteina on ollut kahdenlaisia rooleja: tiedon välittäjä (engl. Mediator of information) ja vaikutuksen levittäjä (engl. Influential spreader) [9]. Näiden lisäksi verkoston staattisuus (engl. Static) eli muuttumattomuus kesken ajon ja dynaamisuus (engl. Dynamic) eli mahdollisuus muutokseen tuovat eri vaikuttamismalleihin sopivat sovelluskohteet.

Solmun keskeisyyden mitta on alun perin esitetty lyhimpien polkujen pituuden keskiarvon käänteislukuna yhdestä solmusta kaikkiin laskettuna [6].

$$C(i) = \frac{N-1}{\sum_{j \neq i} g_{i,j}} \tag{1}$$

Kaavassa $C(i)$ kuvaa keskeisyyttä solmulle i, jossa N vastaa solmujen lukumäärää ja $g_{i,j}$ tarkoittaa geodeettistä etäisyyttä eli lyhimmän polun pituutta solmujen i ja j välillä. Kes-

keisyydellä voidaan mitata verkon globaaleja topologisia rakenteita [30]. Keskeisyys kuvastaa solmun pääsyä muihin solmuihin yhteyksien avulla mahdollisimman nopeasti. Keskeiset solmut levittävät vaikutusta laajemmalle verkostoon nopeammin kuin muut solmut ja täten niitä voidaan pitää tehokkaimpina vaikutuksen levittäjinä [30]. Normalisoitu versio käänteisestä keskeisyydestä muodostaa käsityksen vaikuttajan roolista [58].

$$C_c(i) = \frac{\sum_{j \neq i} g_{i,j}^{-1}}{N-1} \tag{2}$$

Kaavassa $C_c(i)$ on keskeisyyden tunnus. Suurimpana erona Bavelaksen esittämään kaavaan $C(i)$ on summafunktiossa geodeettisten etäisyyksien kääntäminen ennen niiden yhteenlaskemista.

Välittäjyys on mitta solmun tärkeydestä verkostossa ja se normaalisti lasketaan lyhimpien polkujen avulla lähtösolmusta kohdesolmuun. Välittäjyys esiteltiin ensimmäisen kerran Freemanin [27] ja Anthonissen [1] toimesta. Välittäjyys on yleistettävissä solmun vaikuttavuuteen informaation levityksessä verkostossa. Kuitenkin vain lyhimpien polkujen laskeminen tarkoittaisi sitä, että informaatio kulkee vain ja ainoastaan näitä lyhimpiä polkuja pitkin, eikä kaikkia mahdollisia polkuja. Kaikkien polkujen kautta kulkevaa lähestymistapaa esitetään kirjallisuudessa parannuksena edelliseen [11][60]. Välittäjyydessä tarkastellaan solmun i olemista mahdollisimman suurella osalla lyhimmistä poluista solmujen s ja t välillä.

$$C_b(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3}$$

Kaavassa C_b on välittäjyyden tunnus ja solmun i välittäjyyttä mitataan. Merkintä σ_{st} on kaikkien eri lyhimpien polkujen pituuksien summa, jotka kulkevat solmusta s solmuun t ja $\sigma_{st}(i)$ on niiden eri lyhimpien polkujen pituuksien summa, jotka kulkevat tämän välin solmun i kautta. Välittäjyys on määritelty kyseisen solmun kautta levinneen tiedon mitaksi [9]. Välittäjyys voidaan myös normalisoida, jotta lukuarvot olisivat välillä nollasta yhteen kaavan 4 mukaan.

$$C_b'(i) = \frac{C_b(i)}{(N-1)(N-2)/2} = \frac{\sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}}{(N-1)(N-2)/2} \tag{4}$$

Kaavassa $C_b'(i)$ kuvaa normalisoitua välittäjyyttä. Välittäjyyden laskeminen on raskasta johtuen suuresta määrästä polkuja suurissa verkostoissa. Tästä syystä välittäjyyttä kuvaavat matemaattiset mallit mallintavat staattisia eli muuttumattomia verkkoja. Dynaamisia käyttökohteita olisivat muun muassa verkkojen visualisointi ja verkon leviämisrakenteiden syvempi ymmärrys ajan kuluessa [54][55].

Solmujen yhteysaste on kuuluisuutta tai suosiota mittaava tapa. Siinä lasketaan ainoastaan solmun yhteyksien lukumäärä muihin solmuihin tai yksisuuntaisten linkkien tapauk-

sessa, siinä voidaan laskea myös solmuun tulevat linkit. Solmusta lähtevät linkit mittaisivat sen vaikuttavuutta ja solmuun tulevat linkit mittaisivat sen informaation saamista tai vastaanottokykyä. Se antaa karkean arvion solmujen merkittävyydestä, joka on yleisellä tasolla hyvä arvio, mikäli linkkien painoarvoja ei ole tiedossa. Kirjallisuudessa tätä lähestymistapaa ovat käyttäneet muun muassa Sabidussi [69] ja Freeman [28]. Naapurisolmujen yhteysastetta käytetään kirjallisuudessa myös ryhmittymien etsimiseen [76].

2.3 Ryhmittymien etsimisen teoriat

Yhteisöjen etsiminen (engl. Community detection) on noussut tutkittavaksi aiheeksi verkkomallinnuksen ja verkkoanalyysin parissa. Suuressa verkossa olevat pienemmät yhteisöt (engl. Communities) ja ryhmittymät (engl. Clusters) ovat merkittäviä tekijöitä koko verkon topologiasta. Yhteisöt ja ryhmittymät jakavat yhteisiä ominaisuuksia tai jäsenillä on samanlainen rooli suuremmassa verkossa. Yhteiskunta tarjoaa laajan mahdollisuuksien kirjon, josta erilaisia yhteisöjä muodostuu. Tällaisia ovat esimerkiksi kaveripiirit, työpaikka, kaupungit tai Internetin myötä yhteiset ryhmittymät. [25]

Yhteisöjen rakenteiden havainnointi on merkittävää, jotta verkkojen topologian voisi ymmärtää paremmin [76]. Yhteisöjä on etsitty monenlaisilla matemaattisilla tai heuristisilla malleilla [7][17][23][62], mutta vielä kirjallisuudessa ei ole tehokasta ja tarkkaa mallia, jonka avulla voitaisiin suurta monitasoista (engl. Multilayer) verkostoa mallintaa todenmukaisesti. Kaksiosaisen verkon tapauksessa vertailua on tehty algoritmien tehokkuuden osalta ja 10^6 alkiota saavutetaan suuruusluokkana [75]. Verkon muotoa (engl. Shape) tutkiva lähestymistapa on saavuttanut skaalautuvuuden laajemmalle verkostolle, joka sisältää generoituja rakenteellisia tasoja, mutta tasojen tarkastelu aiheuttaa epätarkkuuksia [61].

Limittäisten yhteisöjen (engl. Overlapping communities) löytäminen on edelleen haastava tehtävä eri algoritmeille. Kirjallisuudessa on analysoitu monia ryhmittymien etsimiseen tarkoitettuja algoritmejä ja niiden toiminta on epävarmaa limittäisten yhteisöjen etsimisessä [81]. Kun etsitään limittäisiä yhteisöjä, yksittäinen lineaariseen kompleksisuuteen O(n) skaalautuva ajo ei riitä toteamaan kaikkia yhteisöjä ja ryhmittymiä. Yksittäisissä ajoissa voidaan saavuttaa lineaarinen kompleksisuus [13], mikäli tulokset kattavat vain osan todellisista yhteisöistä. Useimmat limittäisten yhteisöjen tunnistamiseen käytetyt kaavat eivät ole lineaarisesti laskettavissa, kuten Omega [16] tai muut esimerkit Fortunaton oppaassa toteavat [26]. Sisäkkäisiä tai päällekkäisiä yhteisöjä (engl. Nested communities) on nopeampi löytää, sillä tarkasteltava alue pienenee yhteisöjen osalta.

Erilaiset satunnaiskävelyt ovat osoittautuneet hyödyllisiksi kompleksisten verkkojen rakenteellisten ominaisuuksien tutkimisessa [68]. Niiden joukossa on ollut onnistuneesti verkon etsinnässä käytetty polkuaan välttelevä verkko (SAW) (engl. self-avoiding walk), joka vierailee enimmillään vain kerran yksittäisessä solmussa [2]. Perustavanlaatuinen vaihe verkon näkökulmasta on se, että määritellään staattinen verkko saadusta tiedosta,

joka on tehty solmuista ja linkeistä, joihin voidaan soveltaa staattista prosessia kyseisen järjestelmän mallintamiseksi [47]. Tällaisen prosessin jälkeen ryhmittymiä voidaan etsiä erilaisilla menetelmillä, kuten satunnaiskävelyillä, solmujen yhteysasteilla tai yhteisöjen koheesion maksimoinnilla.

2.4 Syrjäytyneisyys

Syrjäytyneisyys (engl. Social exclusion) on ollut ongelmana niin Suomessa kuin monessa muussa Euroopan maassa [4][20][77]. On mielenkiintoista, miten sosialisoituneessa yhteiskunnassa syrjäytyminen on noussut esiin vasta 1990-luvulla EU:n sosiaalipolitiikassa syrjäytymisen vastaisen toiminnan korostettua asiaa [19]. Yleisesti syrjäytyneellä henkilöllä tarkoitetaan sellaista yksilöä, joka on ulkopuolella työnteosta ja opiskelusta, eikä ole saanut peruskoulutusta korkeampaa koulutusta [57]. Toisen määritelmän mukaan syrjäytynyt henkilö voi olla sosiaalisen yhteisön tai ryhmän jäsen, jonka yhteydet ryhmään ovat kadonneet kokonaan tai merkittävältä osin [4][36]. Tällöin useissa tapauksissa haittavaikutuksena syrjäytyneisyydelle on yksinäisyyttä, ahdistusta ja masennusta [5]. Tällaista syrjäytyneisyyttä eli sosiaalista vieraantumista malli arvioi todennäköisyyksien pohjalta.

Eristäytyminen (engl. Social isolation) lasketaan myös monissa tapauksissa syrjäytyneisyyden piiriin, vaikka terminä sillä tarkoitetaan itse aiheutettua eristäytymistä joko koko yhteiskunnasta tai vain lähes kaikista läheisistä ihmisistä. Käytännössä tämä on mahdollista, mikäli henkilö pysyy pitkän ajan kotonaan, haluaa olla vain yksin tai henkilöllä ei ole yhteyksiä toisiin ihmisiin. Eristäytymisen taustalla saattaa olla mielentilahäiriö, jossa esimerkiksi masennuksen takia henkilö vakuuttelee itselleen, että eristäytyminen tuottaa positiivisia tai mukavia kokemuksia [42]. Eristäytymiseen johtavina riskitekijöinä pidetään muun muassa väkivallan uhriksi joutumista, vammaisuutta, yksinasumista ja sosiaalisia vastoinkäymisiä, kuten kiusaamista sekä noloja tai ikäviä tilanteita ryhmässä.

Yhteiskunnassa syrjäytymiseen vaikuttavia tekijöitä on yritetty selvittää ja niitä ovat muun muassa köyhyys, työttömyys, rikostausta, kodittomuus ja perheen hajoaminen [73]. Valtioiden rajoitukset muun muassa uskonnolle [65] ja sukupuolelle [82] syrjäyttää yksilöitä yhteiskunnasta. Yksittäisissä ryhmissä tekijöinä voivat olla lisäksi ihonväri, kasti, varallisuus tai muunlainen erilaisuus [8]. Nämä tekijät vaikuttavat jo itse ryhmään pääsyyn, mutta kulttuurilliset erot ja suvaitsemattomuus voivat aiheuttaa ryhmästä syrjäytymisen myöhemmässäkin vaiheessa, esimerkiksi ryhmädynamiikan muututtua [24].

Yhteisöön kuulumisen mittarina voidaan käyttää lukuarvoa, joka kuvaa eroavaisuutta ryhmän aatteesta

$$Gr_p = |L^{Gr} - L^P| S_{Gr}, \tag{5}$$

jossa yksilön P henkilökohtaista käyttäytymisnormia kuvataan L^P ja tarkastellaan sen eroa ryhmän Gr keskimääräisestä käyttäytymisnormista L^{Gr}. S_{Gr} kuvaa ryhmän sympatiaa erilaisuutta kohtaan. Tätä eroavaisuuden mittaria kuvaa muuttuja Gr_p, jonka suuruus merkitsee ryhmän normeista eroamista, kun taas 0 merkitsee, että yksilön normit vastaavat täysin ryhmän normeja. Liian suuri eroavaisuus ryhmästä johtaa kasvaviin ongelmiin ja lopulta ryhmästä etääntymiseen [49].

Henkilön kuuluessa yhteisöön, hänen norminsa, ajattelumallinsa ja aatteensa saattavat muovautua yhteisön mukaan. Mikäli henkilön aatemaailma ja normit eivät sopeudu yhteisön normeihin ja aatteisiin, se kasvattaa eroavaisuutta henkilön ja yhteisön välillä. Aatepohjaisissa yhteisöissä henkilön aatemaailmalla on eniten painoarvoa yhteisön kannalta, jolloin aatemaailman muut jäsenet eivät välttämättä ole suopeita erilaisuudelle ja voivat tuomita erilaisen henkilön [40].

3. MALLI

Luvussa 3.1 käsitellään tarkemmin mallia, jonka pohjalta sosiaalinen vaikuttaminen lasketaan. Mallin pohjana on käytetty Vesa Kuikan luomaa matemaattista mallia välittäjyydestä ja keskeisyydestä [44] sekä siitä jatkokehitettyä mallia [39]. Tämän mallin ohjelmalliset toteutukset algoritmien osalta käsitellään luvussa 4. Luvussa 3.2 käydään läpi sovellettuja teorioita ryhmittymien muodostuksesta sen matemaattiselta osalta. Ryhmittymien muodostumisen myötä kehittämieni ohjelmien matemaattiset toteutukset käydään myös perusteellisesti läpi.

Tässä tapauksessa syrjäytymisvaarassa olevaa henkilöä tulkitaan ryhmien ja yhteisöjen tahoilta yhteiskunnan sijaan. Syrjäytymisvaarassa olevalla henkilöllä on vähän yhteyksiä toisiin henkilöihin ja hän ei kuulu mihinkään ryhmittymään tai on merkityksettömässä roolissa omassa ryhmittymässään tai laajemmassa yhteisössä. Mallin painottamista ja sen suhdetta syrjäytymiseen käsitellään luvussa 3.3.

Kaikessa mallinnuksessa oleellisena osana on tarkastella mallin luotettavuutta, niin mallinnuksen, mallintajan ja laskennan kannalta. Tässä tapauksessa sitä käsitellään tarkemmin luvussa 3.4, jossa kerrotaan luotettavuuden tekijöistä ja analysoidaan niiden kannalta mallin luotettavuutta. Luvussa arvioidaan tuloksia vaikuttavuuden ja ryhmien muodostamisen osilta muihin kirjallisuudesta löytyviin tuloksiin. Muita käyttökohteita tarkastellaan luvussa 3.5. Päätarkoituksena on antaa laajempaa kuvaa mallin eri osien toimivuudesta osana muunlaista ympäristöä tai eri käyttökohteiden kannalta.

3.1 Sosiaalisen vaikuttavuuden malli

Sosiaalisen vaikuttavuuden malli perustuu eripituisten polkujen kautta kulkeviin todennäköisyyksiin lähtö- ja kohdesolmun välillä. Koko verkon topologia otetaan huomioon, sillä kaikki polut lähtö- ja kohdesolmun välillä otetaan huomioon rajatulla polun pituudella L_{max}. Verkon rakenne on staattinen, sillä se ei muutu kesken yksittäisen ajon. Prosessia mallinnetaan kumulatiivisen Poisson-jakauman avulla, joka kehittyy ajan T mukaan. Myös muita jakaumia voi soveltaa malliin, mutta valittu jakauma on todettu hyväksi ja laaja-alaiseksi, mikäli yksittäiset tapaukset ovat laskennallisesti riippumattomia ja vakiotahdilla eteneviä [49]. Leviäminen on mallinnettu toistuvana prosessina, jossa leviämistä kuvaavat todennäköisyydet lasketaan annetuilla ajanhetkillä [39].

Mallia voi rajoittaa asettamalla maksimaalisen polun pituuden. Käytännössä todennäköisyydet polkujen pituuksilla (>20) ovat merkityksettömiä ja eivät muuta tulosta. Sosiaalisissa verkostoissa, joissa solmuilla on paljon linkkejä, hyviä approksimaatioita saadaan laskettua polun pituudella 10 [39]. Malli perustuu siihen, että kaikki polut lasketaan annettuun maksimipituuteen asti, mikä voi vaikuttaa matemaattisesti hyvinkin raskaalta,

sillä siinä on teoriassa enemmän laskemista kuin lyhimmissä poluissa jokaisesta solmusta jokaiseen. Tätä helpottaa kuitenkin yhteisten polkujen hyödyntäminen sekä mallin että algoritmin puolella [44].

Mikäli polut, joiden pituudet ovat L_1 ja L_2 sisältävät l yhteistä linkkiä polkujensa alussa, voidaan ehdon täyttämät todennäköisyydet laskea $L_1 - l$ ja $L_2 - l$ linkillä. Yhteiset osat polusta huomioidaan laskemalla

$$p_{\{L_1,L_2\}} = p_l(p_{L_1-l} + p_{L_2-l} - p_{L_1-l} * p_{L_2-l}) = \\ p_l p_{L_1-l} + p_l p_{L_2-l} - \frac{p_l p_{L_1-l} * p_l p_{L_2-l}}{p_l} = p_{L_1} + p_{L_2} - \frac{p_{L_1} * p_{L_2}}{p_l}, \quad (6)$$

jossa leviämistodennäköisyyksiä poluista merkitään p_{L_1} ja p_{L_2} sekä yhteisen polun todennäköisyyttä merkitään p_l. Täten $p_{\{L_1,L_2\}}$ on osittainen leviämistodennäköisyys, jossa on yhdistetty polut pituuksilla L_1 ja L_2. Osittaisuus johtuu siitä, että lähes aina yhteisen polun tapauksessa verkostossa on useampi kuin 2 laskettavaa polkua, jolloin laskemalla kaikille näille osille todennäköisyydet, saadaan laskettua kokonaistodennäköisyys.

Painoarvot solmuille ja linkeille on sisällytetty osittaisiin leviämistodennäköisyyksiin p_{L_1}

$$p_{L_1} = p^*_{L_1} \prod_{n \in N_{L_1}} w_n \prod_{y \in Y_{L_1}} w_y, \quad (7)$$

jossa kaikki polun L_1 solmut n merkitään N_{L_1} ja sen linkit eli yhteydet y merkitään Y_{L_1}. Yksittäisiä painoarvoja kuvaavat w_n solmuille ja w_y linkeille ja $p^*_{L_1}$ on ehdon täyttämä todennäköisyys ilman painoarvoja [44].

Kumulatiivinen Poisson-jakauma kuvastaa hetkellistä leviämistä ajanhetkellä T, kun polun pituus L on tiedossa. Poissonin jakaumaa voi käyttäjä muokata sen parametrilla λ, jonka arvon mukaan kertymäfunktio saavuttaa arvon 1 tietyllä ajanhetkellä T. Kaava 8 antaa käytetystä Poissonin jakaumasta matemaattisen kuvan.

$$Po(L) = 1 - \sum_{l=0}^{L-1} e^{-\lambda T} \frac{(\lambda T)^l}{l!}, L > 0; (Po(0) = 1) \quad (8)$$

$$Norm(L) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(L-\mu)^2}{2\sigma^2}}, L \geq 0;\ \mu = 0 \quad (9)$$

$$Tasa(L) = \frac{1}{L_{max}}, L \in [0, L_{max}] \quad (10)$$

$$w\_Tasa(L) = 1 - \frac{L}{L_{max}}, L \in [0, L_{max}] \quad (11)$$

Malliin voi soveltaa muita laskennallisia jakaumia. Ehdotuksena on ollut keskitetty normaalijakauma (Kaava 9), tasajakauma (Kaava 10) ja tasavälinen painotettu jakauma (Kaava 11).

Normaalijakaumassa μ kuvaa keskipistettä, mikä tässä tapauksessa olisi 0 ja σ^2 kuvaa varianssia. Tasajakaumassa ja tasavälisessä jakaumassa L_{max} kuvaa määritettyä maksimipituutta poluille.

Normaalijakauman sovelluskohteet ovat samanlaisia kuin käytetyssä Poissonin jakaumassa, sillä funktioiden ominaisuudet korreloivat keskenään positiivisesti. Sitä voisi soveltaa vaikutuksen leviämiseen. Tasavälistä painotettua jakaumaa kannattaisi käyttää tilanteissa, jossa tapahtumia kuvaa vesiputousmalli, eli aluksi pitää tapahtua ensimmäinen asia, jotta voi tapahtua toinen ja niin edelleen. Silloin ensimmäisen asian tapahtuman todennäköisyys on vahva alussa ja viimeisen tapahtuman todennäköisyys on suurimmillaan lopussa. Tällaista tapahtumaketjua voisi kuvata sähköverkon toiminta, jossa sähkön pitää kulkea ennalta optimoitua reittiä. Pelkkää tasajakaumaa voisi käyttää täysin satunnaisissa tapahtumissa, joissa tapahtumilla ei ole keskenään määrättyä järjestystä. Tällainen tapaus tosielämässä voisi olla eri ihmisten päiväaktiviteetit. Joku ihminen menee ulos aamulla, toinen illalla ja ulkona linnut laulavat satunnaiseen aikaan eri ihmisille. Ihmiset urheilevat, käyvät töissä ja menevät nukkumaan eri aikoina, jolloin sekoittamalla päiväaktiviteetteja ihmisten välillä, saadaan satunnaisia asioita tapahtumaan mallissa.

3.2 Ryhmien muodostamisen malli

Yhteisöjen tunnistaminen on merkittävä työkalu monimutkaisten verkostojen analysointiin, mahdollistaen erikokoisten rakenteiden tutkimisen. Nämä usein liittyvät taustalla olevien verkkojen organisaatioihin ja funktionaalisiin ominaisuuksiin. Yhteisöjen tunnistaminen on osoittautunut arvokkaaksi monilla aloilla, kuten biologiassa, yhteiskuntatieteissä ja sosiaalitieteissä. Huolimatta laajasta mittakaavasta sekä monimutkaisuudesta, yhteisöjä on tutkittu vain vähän sosiaalisessa mediassa [64]. Suurten reaaliverkkojen rakenteiden löytämiseksi tarvitaan algoritmeja, joiden ajoaika kasvaa lineaarisesti $O(N)$. Hierarkkisen ryhmittämisen avulla yhteisöjä voidaan tunnistaa lineaarisella ajalla [3].

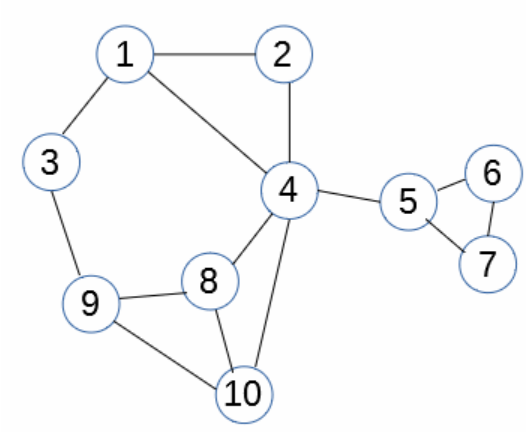
Ryhmien etsiminen ja niiden muodostaminen pohjautuu solmujen läheisyyteen toisten solmujen kanssa. Tähän vaikuttavia tekijöitä yksinkertaisessa verkossa ovat yhteys solmuun ja sen painoarvo sekä polun pituus, mikäli suoraa linkkiä ei ole. Käytetty testidata on julkisista lähteistä saatua painottamatonta dataa, joten ryhmien muodostaminen satunnaisgeneroitujen painoarvojen avulla ei antaisi varmuudella todellista kuvaa, koska painoarvoja ei tiedetä. Painoarvoja ei siis huomioida ollenkaan, tai ne kaikki asetetaan vakioiksi. Painoarvojen vaikutus olisi merkittävä lopputuloksen kannalta, sillä teorian [53] mukaan yhteyksien voimakkuudet määrittävät todennäköisyyden ryhmään jäämiseen tai siitä poistumiseen.

Näiden tietojen pohjalta on kehitetty malli, jossa satunnaisesti valitaan solmu, tarkastetaan ryhmään kuuluvuus ja lisätään solmu sekä naapurit kyseiseen ryhmään, mikäli niillä ei ole ollut aiempaa ryhmää. Tällaista mallia voidaan simuloida miljoonia kertoja minuu-

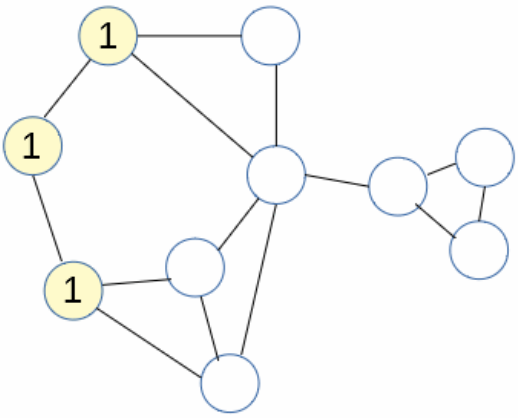
tissa pienillä verkoilla ja suurikin verkko voidaan simuloida kymmeniä kertoja minuutissa. Mallia voidaan pitää satunnaispohjaisena, jossa olemassa olevilla yhteyksillä solmujen välillä on merkitystä.

Malli antaa hyviä approksimaatioita vähemmän vaikuttavista ja ryhmiin kuulumattomista solmuista, kun dataa linkkien painoarvoista ei ole saatavilla. Simulaatioiden lukumäärän nostaminen taas pienentää virhearvojen todennäköisyyttä, jolloin estimoitu virhe jää prosenttiyksiköissä pienemmäksi.

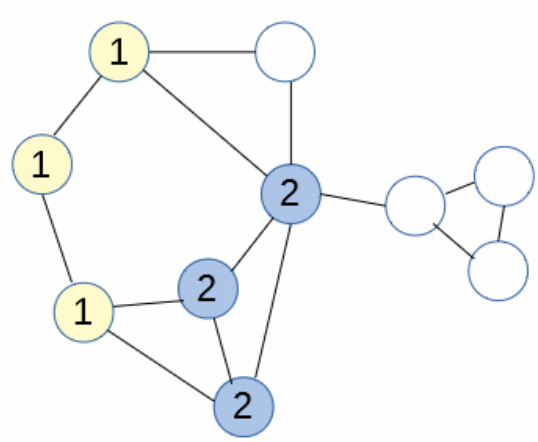
Kuva 1 sisältää 10-solmuinen verkon, johon mallia sovelletaan esimerkkinä. Algoritmi arpoo satunnaislukuja, kunnes kaikki solmut kuuluvat johonkin ryhmään. Ensimmäinen arvottu luku on 3, jolloin ryhmäksi 1 muodostuvat solmut 1, 3 ja 9 (Kuva 2). Seuraava arvottu solmu on 10, joka ei vielä kuulu ryhmään, joten ryhmä 2 muodostuu solmuista 4, 8 ja 10, sillä solmulla 9 on jo ryhmä (Kuva 3). Seuraava arvottu solmu on 1, jolloin ryhmään 1 lisätään solmu 2, sillä solmu 1 kuului jo ryhmään 1 ja muut solmun 1 naapurit kuuluvat jo johonkin ryhmään (Kuva 4). Seuraava arvottu solmu on solmu 8, jonka seurauksena ei tapahdu mitään, sillä solmu ja kaikki sen naapurit kuuluvat jo johonkin ryhmään (Kuva 4). Viimeinen arvottu solmu on 6, jolloin solmut 5, 6 ja 7 muodostavat ryhmän 3. Kaikki solmut kuuluvat johonkin ryhmään, joten yksi simulaatiokierros on suoritettu (Kuva 5). Simulaatioajojen jälkeen ryhmittymille lasketaan todennäköisyydet, eli kuinka usein kyseinen ryhmittymä muodostui (Kuva 6).



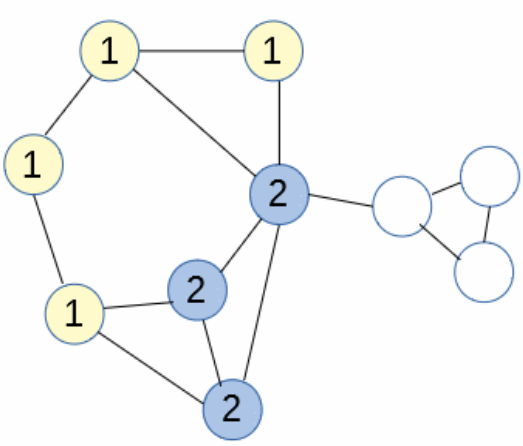
Kuva 1. *Verkon rakenne, solmut numeroitu 1-10*



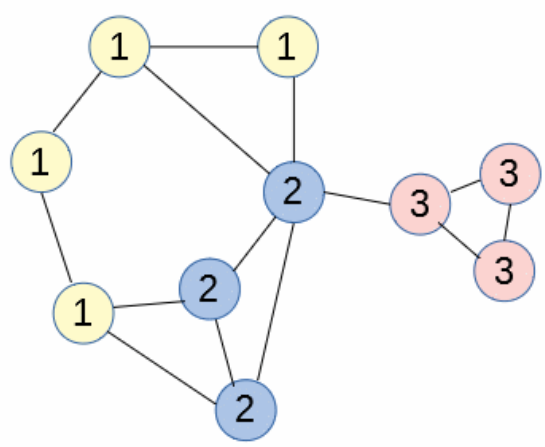
Kuva 2. *Ryhmä 1 arvottu (1,3,9)*



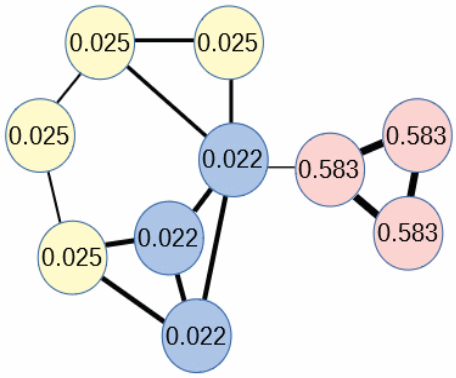
Kuva 3. *Ryhmä 2 arvottu (4,8,10)*



Kuva 4. *Ryhmään 1 lisätty solmu 2*



Kuva 5. *Ryhmä 3 arvottu (5,6,7)*



Kuva 6. *Todennäköisyydet kyseisille ryhmittymille*

Kun solmu 8 arvottiin, mitään ei tapahtunut. Mallin kannalta solmu on hyödytön arpoa, koska tässä tapauksessa sillä ja sen naapureilla on jo ryhmät. Ideaalitapauksessa solmun tulee olla arvottavien listalla vain ja ainoastaan, mikäli sillä tai vähintään yhdellä sen naapurilla ei ole ryhmää. Algoritmillisesti jokaisen arvonnan jälkeen jouduttaisiin tarkastamaan ja päivittämään listaa siitä, mitkä solmut ovat hyödyttömiä arpoa. Tämä kuluttaisi algoritmin osalta aikaa, sillä se on raskas toimenpide. Algoritmin optimoinnissa kerrotaan tästä toteutuksesta ohjelmallisesti lisää.

Yksi simulaatiokierros koostuu esimerkin mukaisesta toimintamallista. Simulaatiokierroksia ajetaan miljoonia, jotta virhe saadaan minimoitua. Jokaisen simulaatiokierroksen jälkeen päivitetään tieto siitä, mitä ryhmiä on muodostunut. Lopuksi tiedot kerätään ja tallennetaan muotoon, josta voidaan tarkastella todennäköisimpiä ryhmäjakoja solmukohtaisesti tai erikseen tarkastella minkä tahansa muodostuneen ryhmän todennäköisyyttä.

Kun tarkastellaan suurimpia todennäköisyyksiä tietyille minimiryhmäkoille, voidaan täten arvioida myös limittäisiä ryhmiä. Esimerkiksi hollantilaisopiskelijoiden verkostossa havaitaan, että erikokoiset ryhmät muodostavat jakoja eri lailla ja limittäisiä ryhmiä muodostuu. Pienikokoisilla ryhmillä yksittäinen solmu saattaa kuulua kahteen tai kolmeen ryhmään, mutta suurikokoisemmilla ryhmillä pieni ryhmä saattaa olla osana monia suurempia ryhmiä. Pienikokoinen ryhmä saattaa myös jakaantua kahteen eri suureen ryhmään, mikäli pienikokoinen ryhmä ei ole tiivis ja sen jäsenillä on yhteyksiä molempiin suurempien ryhmien jäseniin. Ryhmien jaottelu vaatii lisätutkimusta.

Toinen ryhmien etsimiseen käytetty algoritmi perustuu lokaalien maksimien etsimiseen, jossa satunnaisesti generoimalla ryhmä jaetaan kahtia, sen jälkeen vertaillaan ensimmäisen ja toisen ryhmän solmuja keskenään, laskemalla koheesiot ryhmille 1 ja 2 (Kaava 12) mukaan ja summaamalla ne (Kaava 13) mukaan. Mikäli lokaali maksimi löytyy, eli kummastakaan ryhmästä ei voi siirtää solmua toiseen, jotta koheesioiden summa kasvaisi (Kaava 14) mukaan, yksittäinen ryhmäjako on löytynyt. Tämän jälkeen voidaan tarkastella muita jakoja tai jakaa löydettyjä ryhmiä entisestään algoritmin avulla.

$$Coh(G) = \frac{1}{N_G}\sum_{n=1}^{N_G} C_b(n),\ n \in \{G\} \tag{12}$$

$$Coh_i(G_1, G_2) = Coh(G_1) + Coh(G_2) \tag{13}$$

$$\Delta Coh_i(G_1, G_2) = Coh(G_1 - \{n_1\}) + Coh(G_2 + \{n_1\}) - Coh_i(G_1, G_2),\ n_1 \to G_2 \tag{14}$$

$$\Delta Coh_i(G_1, G_2) = Coh(G_1 - \{n_1\} + \{n_2\}) + Coh(G_2 - \{n_2\} + \{n_1\}) - Coh_i(G_1, G_2) = Coh(n_1 - n_2) * (-Coh(G_1 - \{n_1\}) + Coh(G_2 - \{n_2\})),\ n_1 \to G_2 \,||\, n_2 \to G_1 \tag{15}$$

Koheesion Coh laskemisessa ryhmälle G termi n kuvaa solmua, joka kuuluu ryhmään G. Ryhmän G jäsenten määrää kuvaa N_G ja koheesiota lasketaan keskeisyyden mitan C_b avulla. Koheesioindeksiä Coh_i lasketaan ryhmille G_1 ja G_2. Muunnos summien laskemisessa tapahtuu poistamalla solmu n_1 tai n_2 toisesta ryhmästä ja lisäämällä ne toiseen (Kaava 14). Laskemista nopeuttaa se, että koheesioindeksiä ei tarvitse laskea joka kohdassa uudelleen, vaan voidaan vertailla muutosta. Kaava 15 kuvaa koheesion muutosta, jossa solmu n_1 siirretään ryhmään G_2 ja solmu n_2 siirretään ryhmään G_1. Muutoksessa lasketaan molemmille solmuille lukuarvot, joista negatiivinen osuus tulee vanhasta ryhmästä poistamisesta ja positiivinen osuus uuteen ryhmään lisäämisestä.

3.3 Mallien painottaminen

Vaikuttavuuden malli on todettu toimivaksi ja antaa omaan aihealueeseensa hyviä tuloksia. Sen avulla ei yksin voida arvioida syrjäytyneisyyttä verkoston rakenteen mukaan. Verkon rakenteen mukaisia yhteisöjä kuvaamaan on luotu algoritmi, joka simuloi erilaisia

ajoja verkosta. Näiden avulla saadaan todennäköisyydet eri ryhmille sekä solmun yksinäisyydelle. Yhdistämällä nämä kaksi mallia, saadaan jaettua mallien merkittävyyttä, jotta kokonaisuus solmujen syrjäytymisestä olisi todenmukaisempi.

Mallien yhteistoiminta muodostaa syrjäytymisvaarassa oleville solmuille tarkemman arvion. Vaikuttamismalli osoittaa sen, kuinka suuri vaikutusvalta yksilöllä on yhteisössä tai koko verkossa. Ryhmittymien etsimiseen käytetty malli tarkastelee yksilön sijaintia ryhmittymissä tai yhteisöissä. Näistä molemmista saadaan arvokasta tietoa yksilöstä, jonka perusteella on helpompi tehdä riskianalyysiä yksilön tilasta.

Kokonaismallin tuottamaa laskentamallia voidaan evaluoida ryhmäytymisen mukaan (Kaava 16), jossa $Gr_i(n)$ kuvaa ryhmään kuulumisen indeksiä, S kuvaa alaindeksin ryhmäkokoa, p alaindeksin todennäköisyyttä ja N solmujen lukumäärää. Alaindeksi $Gmax$ kuvaa suurimman todennäköisyyden ryhmää, jossa solmuja on vähintään kaksi. Alaindeksi n kuvaa ”ryhmää”, jossa solmu on yksin. $P_{s,t}$ kuvaa leviämistodennäköisyyttä solmusta s solmuun t.

$$Gr_i(n) = \frac{1+\frac{\sqrt{S_{Gmax}}p_{Gmax}-\sqrt{S_n}p_n}{\sqrt{S_{Gmax}}p_{Gmax}+\sqrt{S_n}p_n}+\frac{k}{N}\sum_{t=1}^{N}P_{s,t}}{2+k} \tag{16}$$

Tämän avulla saadaan laskettua syrjäytymisindeksi $O_i(n)$ (Kaava 17) avulla. Syrjäytymisindeksin on tarkoitus rajata selkeästi, mitkä solmuista ovat mahdollisesti syrjäytyneitä.

$$O_i(n) = \begin{cases} 0, & G_i(n) > 0.5 \\ 1 - Gr_i(n) * 2, & G_i(n) \leq 0.5 \end{cases} \tag{17}$$

Näissä kaavoissa mallien keskinäistä tasapainoa voi painottaa muuttujan k avulla. Mikäli halutaan vertailla ainoastaan ryhmittymismallin tuloksia, annetaan k:lle arvoksi 0. Mikäli halutaan painottaa vaikuttamismallia, voidaan k:lle antaa suuri kokonaisluku. Suositeltu tasapaino saadaan asettamalla k:n arvo välille yhdestä neljään, jolloin molempien mallien painoarvo kokonaisuudesta on välillä 33 ja 67 prosenttiyksikköä.

Kaavat 16 ja 17 on kehitetty kyseistä mallia varten. Syrjäytymisen tarkastelussa keskityttiin henkilön vaikutusvaltaan koko yhteisössä sekä henkilön todennäköisiin ryhmiin ja niiden kokoihin. Henkilön ollessa täysin eristyksissä muusta yhteisöstä, tulisi todennäköisyyden olla 100 prosenttiyksikköä. Mikäli henkilö on vaikutusvaltaisessa asemassa yhteisössä ja ei esiinny ryhmittymissä yksin, ei hänellä ole riskiä olla syrjäytynyt mallin mukaan. Tosielämässä syrjäytymistä määrittävät lukuisat muut tekijät, mutta mallin on tarkoitus antaa approksimaatio syrjäytyneisyydestä todennäköisyyden keinoin tai sanallisesti: ”syrjäytynyt” eli $O_i(n) \geq 0.5$, ”mahdollisesti syrjäytynyt” eli $0.5 > O_i(n) > 0$, ”ei syrjäytynyt” eli $O_i(n) = 0$.

3.4 Mallin luotettavuus

Mallin luotettavuuden arviointi on tärkeä osa kokonaisuutta, sillä epäluotettava malli ei ole hyödyllinen. Tieteellisessä mielessä siitä voi olla haittaa, mikäli sitä pidetään hyvänä mallina. Mallin luotettavuutta kuvataan tieteellisin perustein. Arviointitaulukko on kirjoittajan näkemys mallin luotettavuudesta. Mallintaminen itsessään koostuu monista osa-alueista ja siihen vaaditaan tietotaitoa, jotta toimiva malli saadaan aikaiseksi.

Esimerkkitulokset generoidaan mallilla julkisista lähteistä saaduista verkoista. Julkisesti saatavilla olevat verkot on otettu käyttöön, koska saman alan tutkimuksissa on myös käytetty kyseisiä verkkoja. Verkkojen oikeellisuus ja luotettavuus tulee kuitenkin kyseenalaistaa, sillä sosiaalisten verkostojen datan kerääminen on haastavaa [10]. Verkkojen käyttäminen eri tutkimuksissa lisää kuitenkin luotettavuutta kyseisiä verkkoja kohtaan. Mikäli kirjallisuudessa on esitetty tuloksia, jotka ovat lähes tai täysin samanlaiset kuin luodun mallin tapauksessa, voi mallia pitää luotettavampana.

Taulukko 1. *Mallintamisen tulosten luotettavuuteen vaikuttavia tekijöitä [66]*

Ominaisuus/tekijä	**Vaikutus tuloksiin**
Ohjelmiston taso	Mitä koetellumpi ja kehittyneempi ohjelmisto, sitä parempia ja luotettavampia tuloksia voidaan odottaa.
Mallintajan asiantuntemus	Puutteellinen teoreettinen tietämys aiheeseen liittyvästä fysiikasta ja laitteiden ominaisuuksista ja teknisistä reunaehdoista voi johtaa vakaviin virhepäätelmiin.
Lähtötiedot, täsmällisyys ja luotettavuus	Jos lähtötiedot ovat puutteelliset, ei malli vastaa todellista järjestelmää täysin ja tuloksien luotettavuus heikkenee.
Toiminnallisen suunnittelun tehtävänannon selkeys	Tehtävänannon tulee olla riittävän selkeä, jotta toiminnallisen suunnittelun lopputulos vastaa tavoitetta.
Suunnitelman ja tehtävänannon yhdenmukaisuus	Jos suunnitelma ei vastaa tehtävänantoa, eivät tuloksetkaan vastaa tavoitetta.
Mallista käytettyjen teoreettisten komponenttien ja lopullisten asennettujen komponenttien ominaisuuksien yhdenmukaisuus	Jos mallintamisessa ei ole käytettävissä lopullisten komponenttien tietoja, voidaan harkiten käyttää tilastollisia tietoja. Tällöin on kuitenkin syytä mainita, että tiedot ovat alustavia ja että analyysi on syytä päivittää lopullisten tietojen mukaan.
Fysikaaliset poikkeamat lähtötiedoissa	Esim. vedessä oleva ilma

Muita luotettavuuteen vaikuttavia tekijöitä ovat muun muassa ohjelmiston taso, mallintajan asiantuntemus ja lähtötiedot [66]. Nämä tekijät ja niiden vaikutukset tuloksiin on esitelty taulukossa 1.

Taulukko 2. *Luotettavuuden arviointitaulukko tekijän näkökulmasta*

Ominaisuus	**Arvio**
Ohjelmiston taso	Vaikuttamisessa käytetty malli on tasoltaan vähintään hyvä. Se on tehokas ja siitä tuleva virhemarginaali on käyttäjän määriteltävissä. Ryhmittymien etsimisessä kokeiltiin erilaisia lähestymistapoja. Näistä valittiin sopivin syrjäytyneisyyden kannalta. Tasoltaan tämä on kohtalainen.
Mallintajan kokemus ja tietotaito	Mallintajalla on kokemusta ohjelmoinnista yli 5 vuotta ja mallintamisesta n. 1 vuosi. Tämä on selkeä riski luotettavuuteen. Tosin, vaikuttamisessa käytetyssä mallissa suunnittelijana toimineena henkilöllä on pitkä kokemus matematiikasta.
Testidatan oikeellisuus ja luotettavuus	Testidataa suurille verkoille ei ole paljoa ja niiden vertailua todellisuuden kanssa ei ole tehty. Pienillä verkoilla on esimerkkinä käytetyssä tapauksessa tarkka testidata ja todellisuuden tapahtumat tukena oikeellisuudelle. Tieteellisellä tasolla tämä on keskiarvoa hieman parempi.
Testidatan kvantitatiivisuus	Testidataa eri verkoista on sovellettu tähän malliin kohtuullisesti. Useita testidatoja, jotka ovat vapaasti saatavissa ja tieteellisesti käytettyjä, käytetään myös tässä tutkimuksessa. Luotettavuuden kannalta tämä on hyvä.
Mallin teoriapohja	Erilaisiin lähestymistapoihin on perehdytty ja mallille (etenkin vaikuttamismallille) on tehty matemaattisesti pätevä malli asiantuntijan toimesta. Tämä on erinomaisella tasolla.
Mallin sovellettavuus muuhun käyttötarkoitukseen	Malli koostuu osista, jotka soveltuvat eri tieteenaloille ja useaan käyttötarkoitukseen. Kuitenkin kokonaismalli ei ole suoraan sovellettavissa muihin kuin sosiaalisiin verkostoihin. Tämä on tavallaan hyvä, ettei malli yritä olla jokaisella osa-alueella kohtalainen vaan ennemmin yhdellä osa-alueella erinomainen.
Algoritmin toiminta ja tehokkuus	Vaikuttamismallin algoritmi on nopea ja tehokas. Ryhmittymien etsimisessä käytettävä algoritmi soveltuu myös suuremmille verkoille,

	mutta sen tehokkuus ei yllä vaikuttamismallin tasolle. Luotettavuuden kannalta nämä ovat kohtalaisia tai hyviä.
Tulosten samankaltaisuus	Tulokset vaikuttamismallin osalta vastaavat muita kirjallisuudessa esitettyjä tuloksia. Ryhmittymien etsimisessä käytettävää mallia ei ole vertailtu. Kokonaisuutena tämä on kohtalaisella tasolla.

Luotettavuuden arviointitaulukko (Taulukko 2) osoittaa, että usealla osa-alueella on puutteita, mutta tulosten kannalta mallia kannattaisi kokeilla muille verkoille. Suurimpana ongelmana on vähäinen aiheen matemaattinen käsittely kirjallisuudessa, jolloin vertailukohteita ei ole kokonaismallille. Todellisuutta kuvaavan mallin luominen on vaikeaa [15]. Mallin yksittäisiä osia pystytään vertailemaan ja ne ovat yhteensopivia keskenään, joten niistä saa jonkinlaisen kuvan mallin luotettavuudesta. Mikäli ohjelma tulee julkiseen käyttöön, voisi sen toimintaa arvioida tarkemmin, koska käytössä olisi enemmän resursseja.

3.5 Mallin käyttökohteet ja juridiset esteet

Malli on suunniteltu ensisijaisesti sosiaalisesti syrjäytyneiden löytämiseksi verkostosta, jonka rakenne tiedetään. Mallia voidaan soveltaa verkostoihin, joiden koko rakennetta tai painoarvoja ei tiedetä. Muita käyttökohteita on listattuna kokonaismallille sekä sen osille, joita on jo käytetty tai käytetään myös muihin tarkoituksiin.

Mallia on testattu tapauksissa, jotka käsittävät pieniä yhteisöjä ja laajempia verkostoja. Mallin mahdollisena käyttökohteena ovat syrjäytyvien tai yksinäisten lasten löytäminen ryhmästä, esimerkiksi luokasta tai laajemmin koko koulusta. Tämän lisäksi mallin tulisi antaa approksimaatioita työpaikan tai harrastuksen ihmissuhteita kuvaavista verkoista. Suurista verkostoista (>1000 solmua) syrjäytyneen tunnistaminen voi olla vaikeaa, mutta luvussa 5.2 arvioidaan sen toimivuutta vähintään tuhannen solmun verkoissa.

Malli koostuu kolmesta suuremmasta osasta, joista ryhmittymien etsimistä ja vaikutuksen leviämistä voi käyttää erikseen omissa käyttötarkoituksissaan. Vaikutuksen leviämistä mittaava malli on todettu hyväksi sosiaalisten verkostojen mallinnuksessa, jossa ihmisten välisiä suhteita analysoidaan vaikutusvallan ja merkittävyyden keinoin. Sen avulla voidaan myös tarkastella välittäjyyttä tiedon siirtämisessä tai ylipäätään asian viestimisessä paikasta A paikkaan B tai kaikille vastaanottajille. Ryhmittymien etsiminen on tehty uutena toteutuksena ja sen käyttökohteita oletettavasti on erilaisten ryhmien, järjestöjen tai yhteisöjen etsiminen laajemmasta painottamattomasta verkostosta. Sitä voisi myös soveltaa ajatuskarttojen (engl. Mind map) lokeroimiseen tai pienempiin osiin jakamiseen. Tätä ei ole testattu, mutta tämä voisi auttaa osakokonaisuuksien ymmärtämisessä.

Tämän mallin osia voisi käyttää toisen yhteiskunnallisen ongelman ratkaisuna: rasismin ennaltaehkäisyssä. Mikäli malliin liitetään tuki solmun attribuuteille (engl. Node attributes), algoritmien avulla joukosta ihmisiä olisi mahdollista muodostaa ryhmiä, joissa erilaisia kulttuuritaustoja yhdistetään samaan ryhmään mahdollisten rasistien kanssa. Malliin voitaisiin soveltaa tiimiyttämismetodia [45], jotta ryhmädynamiikka paranisi kansan keskuudessa [50]. Rasismi ei aina ole tiedostettua, sillä monilla ihmisillä on ennakkoluuloja (engl. Bias) toisia etnisiä ryhmiä kohtaan vähäisen kanssakäymisen takia [74].

Algoritmin soveltaminen voi olla vaikeaa syrjäytymisen ennaltaehkäisemiseksi, sillä datan keräämisessä suoraan lapsilta on esteitä huoltajuuden myötä Suomen säädöskokoelmassa (SDK 8.4.1983/361 1:4§). Huoltajan tulee antaa lupa tietojen keräämiseen esimerkiksi kouluympäristössä. Lisäksi tietokannan ylläpito automaattisesti velvoittaa ilmoittamaan asiasta tietosuojavaltuutetulle (SDK 22.4.1999/523 8:36§). Yleisesti tietoja saa käsitellä henkilön suostumuksella (SDK 22.4.1999/523 2:8§), tutkimus- (SDK 22.4.1999/523 4:14§) tai tilastotarkoituksessa (SDK 22.4.1999/523 4:15§) sekä henkilömatrikkelina (SDK 22.4.1999/523 4:17§).

Rasismin ennaltaehkäisemiseksi ongelmaksi tulee arkaluonteisten tietojen käsittelykielto (SDK 22.4.1999/523 3:11§), sillä tietokannassa olisi oleellista olla etninen alkuperä tai muu arkaluonteinen tieto, jotta suvaitsevaisuutta pystytään lisäämään henkilöiden välillä. Tässäkin tapauksessa kaikilta osallisilta tulee olla lupa tietojen käyttöön (SDK 22.4.1999/523 3:12.1§) tai tietosuojalautakunnalta vedoten yleiseen etuun (SDK 22.4.1999/523 9:43.2§).

4. ALGORITMI

Luvussa 4.1 käydään läpi vaikuttamismallin algoritminen näkökanta ja tarkastellaan sen skaalautuvuutta. Luvussa 4.2 kerrotaan ryhmittymien etsimiseen käytetyn algoritmin ohjelmallisesta toteutuksesta ja siinä käytettävän simuloinnin skaalautuvuudesta. Luvussa 4.3 selitetään perusteellisesti mallit yhdistävän algoritmin toteutus ohjelmallisesti sekä perustellaan kaavan toimintaa syrjäytymisen mittana.

4.1 Vaikuttamismallin toteutus

Vaikuttamismalli ottaa huomioon kaikki polut käyttäjän rajaamalla pituudella, eikä vain lyhimpiä polkuja solmusta a solmuun b. Polun pituuden vaikutus näkyy Poissonin kumulatiivisessa mallissa, jossa solmulle annetaan pienempi painoarvo polun pituuden kasvaessa. Vaikuttamismalli perustuu esitettyyn algoritmiin [39] ja malliin [44]. Vaikuttamismalli on toteutettu C++-ohjelmointikielellä. Tuloksien analysoimisessa on muun muassa käytetty Python-toteutuksia. Poisson-toteutus on esitetty alla (Ohjelma 1).

```
#include <cmath>
#include <boost/math/distributions/poisson.hpp>
#include <boost/math/distributions/complement.hpp>
#include <limits>

extern "C" double poisson(int num, double t, double lambda)
{
    if(num == 0) {
        return 1.0;
    }
    boost::math::poisson_distribution<> p(lambda*t);
    auto val = cdf(complement(p, num-1));
    auto epsilon = std::numeric_limits<double>::min();
    if( val < epsilon ) {
        return epsilon;
    }else {
        return val;
    }
}
```

Ohjelma 1. *Poisson-toteutus*

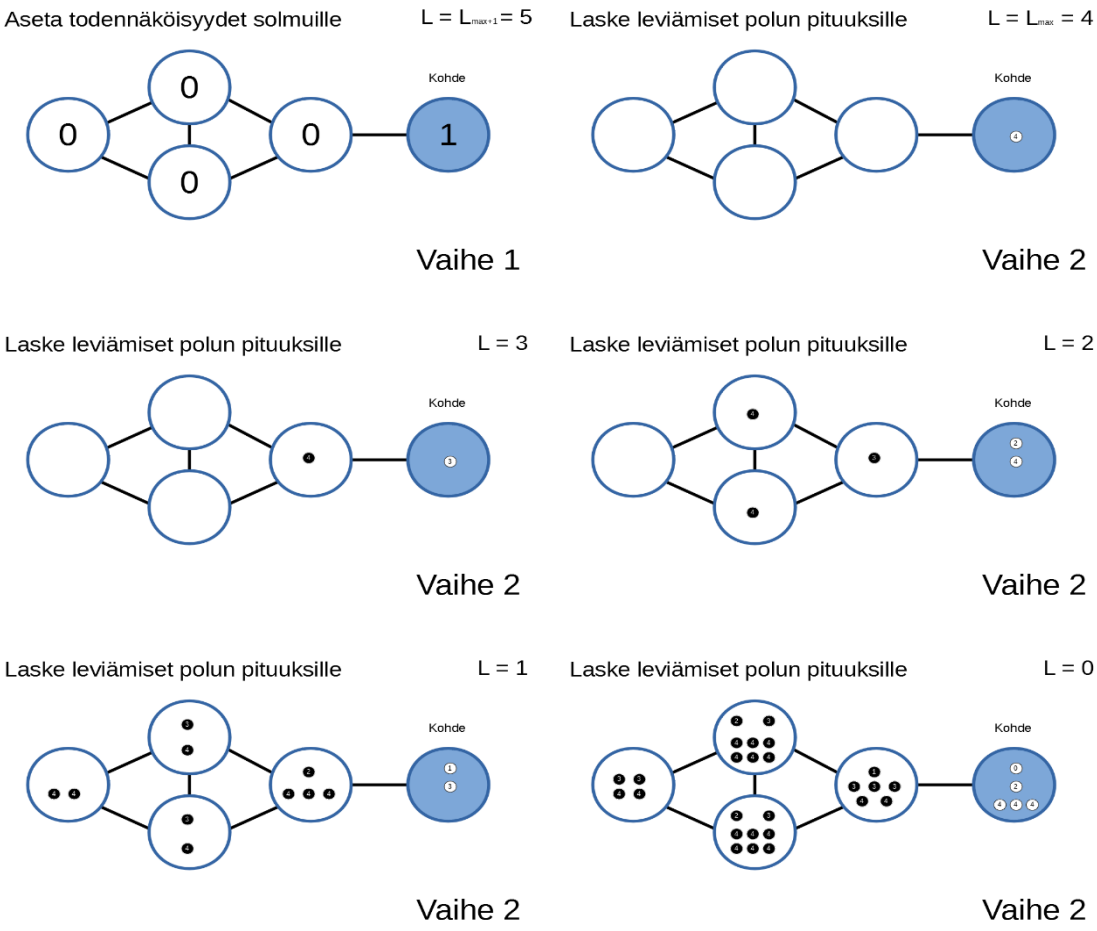
Algoritmi käy polkuja läpi leviämisaskel kerrallaan, jotta kaikkia polkuja ei tarvitse käydä erikseen läpi. Tällöin algoritmi prosessoi samanlaiset kohdat verkoston puurakenteesta (engl. Traversal tree) ainoastaan kerran. Laskeakseen leviämistodennäköisyydet, algoritmin tarvitsee yhdistää todennäköisyydet kaikilta polun pituuksilta, jotka ovat alle valitun maksimaalisen polun pituuden L_{max}. Algoritmi käy läpi polun pituuksia käänteisessä järjestyksessä aloittaen asettamalla $P_{s,t,L_{max}+1} = 0$, sillä maksimi polun pituutta L_{max} suu-

remmat arvot ylittävät tarkastelualueen ja suurilla L_{max} arvoilla leviämistodennäköisyydet eivät enää muutu. Sen jälkeen algoritmi vähentää polun pituutta yksi kerrallaan arvosta L_{max} arvoon 0 [39]. Algoritmin toteutus on näkyvissä kohdassa (Liite A).

Polun pituuden L arvoilla lasketaan osittaisia leviämistodennäköisyyksiä (Kaava 18) mukaan, jossa i kuvaa solmua, johon leviäminen tapahtuu naapurisolmusta j. Tämän laskemista helpottaa se, että kaikkien naapurisolmujen todennäköisyydet tiedetään pituudella $L + 1$.

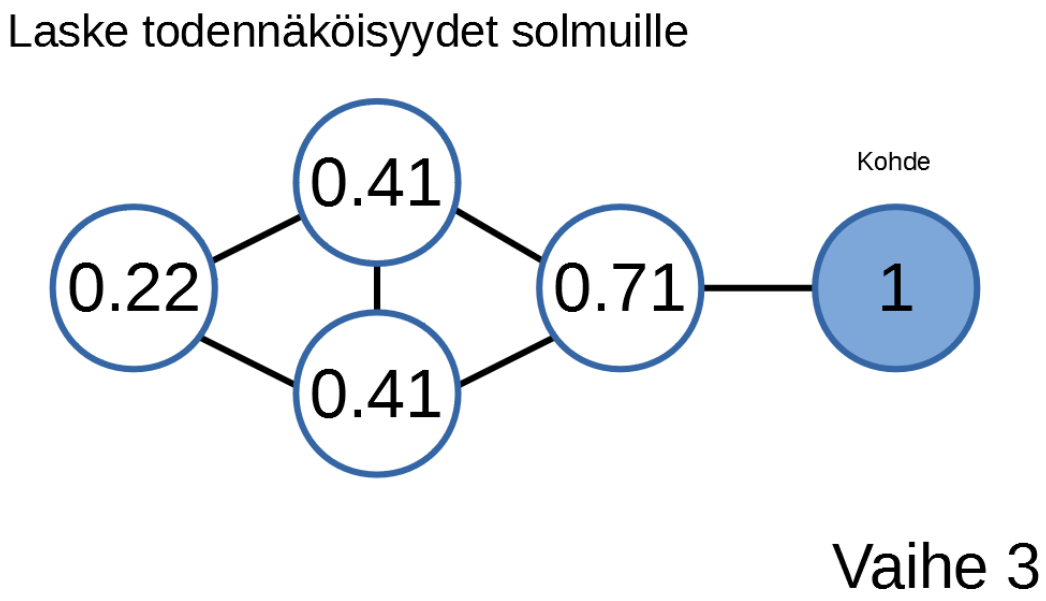
$$P^{*}_{i,t,L} = P_{i,t,L} + P_{j,t,L+1} - \frac{P_{i,t,L} \cdot P_{j,t,L+1}}{Po(L)} \tag{18}$$

Kun algoritmin laskenta pääsee polun pituuteen 0, saavutetaan tilanne, jossa $P_{s,0} = P_{s,t}$ eli kertynyt todennäköisyys kaikille lasketuille polun pituuksille saavuttaa koko todennäköisyyden lähtösolmusta s kohdesolmuun t. Laskettaessa tämä kaikille solmuille, saadaan kaikista solmuista todennäköisyydet yhteen solmuun ja päinvastoin. Kuvasarja (Kuva 7) esittää leviämistä polun pituuksille $L_{max} = 4$.



Kuva 7. *Leviämistapahtumat 5-solmuisessa verkossa polun pituudella* $L_{max} = 4$

Tässä tapauksessa vaihe 1 kuvaa yksittäisen kohdesolmun valitsemista, jolle asetetaan todennäköisyydeksi yksi, muille solmuille leviämistodennäköisyydet asetetaan nollaan. Vaihe 2 kuvaa leviämistapahtumien mallinnusta verkossa, jossa polun pituus L alkaa arvosta L_{max} ja loppuu arvoon 0. Vaiheessa 2 jokaiselle polun pituudelle lasketaan väliarvot todennäköisyyksille, käyttäen naapurisolmujen välisiä painoarvoja.



Kuva 8. *Leviämisten myötä lasketut todennäköisyydet esimerkkiverkon solmuille*

Tämän jälkeen siirrytään vaiheeseen 3, jossa leviämisten myötä lasketaan todennäköisyydet edellisen vaiheen $L = 0$ mukaan ja niistä saadaan lopulliset todennäköisyydet yksittäiselle kohdesolmulle. Vaihetta 3 selventää (Kuva 8), joka on laskettu arvoilla $L_{max} = 20, T = 1.0$.

4.2 Ryhmittymien etsiminen verkostosta

Painotetuissa verkostoissa eri solmujen väliset linkkien painoarvot ovat merkittävässä osassa verkon solmujen vaikuttavuuteen. Painottamattomassa verkostossa verkon topologia, naapurisolmut ja yhteysaste ovat päätekijät vaikuttavuuden mittaamiseksi. Ryhmien etsiminen verkostosta pohjautuu samoihin tekijöihin.

Aikaisempi toteutus on tehty Python-ohjelmointikielellä, jossa lasketaan koheesiota ryhmille. Päätarkoituksena on löytää koheesioindeksin lokaaleja maksimeja verkostoista. Lokaalin maksimin tapauksessa, mitään solmua siirtämällä, ei saada suurempaa arvoa koheesioindeksille. Tällainen ryhmäjako tallennetaan ja siitä lasketaan koheesioindeksi (Ohjelma 2) mukaisesti. Lähestymistapana tämä sopii sellaisille verkostoille, joissa painoarvot ovat annettu tai asetettu vakioiksi. Algoritmi antaa tarkkoja tuloksia perustuen vaikuttavuuteen, mutta on O-notaation kannalta hidas laskettava. Pahimmillaan

$O((n/2)\text{^}3 * S)$, jossa n on solmujen määrä ja S simulaatioiden määrä. Lisäksi satunnaisuus ryhmien valinnassa voi johtaa samankaltaisiin simulaatioihin, jolloin tehokkuus kärsii.

```
def compute_cohesion_index(first, second):
    v_a = 0.
    v_b = 0.
    group1 = list(itertools.combinations(first, 2))
    group2 = list(itertools.combinations(second, 2))
    for item in group1:
        v_a += v_i(item[0], item[1])
    for item in group2:
        v_b += v_i(item[0], item[1])
    return v_a+v_b
```

Ohjelma 2. *Koheesioindeksin laskenta valituille ryhmille*

Toinen toteutus on tehty myös Python-ohjelmointikielellä, jossa käyttäjä saa valita minkä kokoisia ryhmiä tarkastellaan. Algoritmi kirjoittaa tiedostoon valitun kokoisia ryhmiä halutun määrän jokaiselle solmulle. Käyttäjä voi tarkastella solmukohtaisesti todennäköisyyksiä eri ryhmien muodostumiselle. Algoritmin simulaatiovaihetta näyttää (Ohjelma 3) ja solmukohtaista todennäköisyyttä järjestävää algoritmia esittää (Ohjelma 4). O-notaatiolla kuvattuna algoritmi on kohtalaisen tehokas $O(n\text{^}2 * S)$, mutta epävarmuustekijänä on satunnaisuus ryhmien muodostumisessa. Simulaatioiden lukumäärä vähentää yksittäisen virheen suuruutta, sillä todennäköisyys suurelle poikkeamalle pienenee [63].

```
def driveSimulation(args, dict, nodes, tot):
	groups = {}
	for simulation in range(args.simulations):
		print("Simulation round: ", str(simulation), "/", str(args.simulations), "\r",
		end="")
		_groups = []
		_nodesrem = list(nodes)
		_ncnodes = list(nodes)
		while len(_nodesrem) != 0:
			_newn = random.choice(_ncnodes)
			_newn_in_cluster = False
			_newn_group_index = -1

			for _ind, _cluster in enumerate(_groups):
				if _newn in _cluster:
					_newn_in_cluster = True
					_newn_group_index = _ind
					break
			if not _newn_in_cluster:
				_groups.append([_newn])
				_nodesrem.remove(_newn)
				_ncnodes.remove(_newn)
			for _neighbour in dict[_newn]:
				if _neighbour in _nodesrem:
					_groups[_newn_group_index].append(_neighbour)
					_nodesrem.remove(_neighbour)
		for x in _groups:
			_new_x = sorted(x)
			_new_str = " ".join(str(_x) for _x in _new_x)
			if _new_str in groups:
				groups[_new_str] += 1
			else:
				groups[_new_str] = 1
```

Ohjelma 3. *Simulaatioita ajava ohjelma, jossa tarkastellaan ryhmiä*

```
for i in range(-1, -1-args.results_per_g, -1): # -1, -2, ..., -10.
	if groups[_g] > top_perc[_sub][i]:
		if i < -1:
			top_perc[_sub][i+1] = top_perc[_sub][i]
			top_grp[_sub][i+1] = top_grp[_sub][i]
		top_perc[_sub][i] = groups[_g]
		top_grp[_sub][i] = _g
	else:
		break
```

Ohjelma 4. *Solmukohtainen järjestämisalgoritmi todennäköisyyksien mukaan*

Ryhmittymien etsiminen perustuu solmun yhteyksiin. Koko verkostosta arvotaan solmu, joka asettaa itselleen ja naapureilleen ryhmän. Mikäli valittu solmu kuuluu jo ryhmään, sen naapurisolmut, jotka eivät kuulu mihinkään ryhmään, asetetaan kuulumaan samaan ryhmään kuin solmu itse. Simulointia jatketaan, kunnes jokainen verkon solmu kuuluu johonkin ryhmään. Simulaatioita suoritetaan käyttäjän valitsema määrä. Todennäköisyydet muodostuvat kaavan avulla, jossa saman ryhmän esiintymisten lukumäärä jaetaan simulaatioiden lukumäärällä. Esimerkiksi ryhmää (Kuva 5), joka koostuu solmuista 5, 6 ja 7 esiintyy simulaatiossa 582 987 kertaa ja simulaatioiden kokonaislukumäärä on miljoona kappaletta. Silloin todennäköisyys ryhmän (5, 6, 7) muodostumiselle on noin 58,3 prosenttia (Kuva 6).

Toisessa toteutuksessa on mahdollistettu yksittäisten solmujen esiintyminen ryhmänä, jotta syrjäytyneitä voidaan tarkastella todennäköisyyksien nojalla. Käyttäjä pystyy valitsemaan tarkasteltujen ryhmittymien koon, simulaatioiden lukumäärän sekä solmukohtaisten tulosten lukumäärän. Tulosten analysoiminen on helpompaa ja nopeampaa, sillä algoritmi järjestää ryhmäkohtaiset todennäköisyydet suuruusjärjestykseen solmu kerrallaan. Tulokset tallennetaan tiedostoon jatkokäsittelyä varten.

4.3 Mallit yhdistävä algoritmi

Mallit yhdistävä algoritmi on käytännössä Pythonilla tehty skriptitiedosto, joka lukee tiedot kaikille muuttujille, joita kaavassa (15) käytetään. Kun tiedot on koottu, laskee algoritmi kyseisen kaavan avulla kaikille solmuille indeksin ryhmään kuulumiselle $G_i(n)$ tai syrjäytymisindeksin $O_i(n)$, käyttäjän antaman painoarvon k mukaan. Lopuksi ohjelma tallentaa ajotiedoston tekstimuotoon, mikäli ajon parametreja halutaan tarkastella. Ohjelma tallentaa myös pelkästään numeroarvoja sisältävän tiedoston, jonka saa helposti siirrettyä taulukko-ohjelmaan tulosten visualisoimiseksi.

```
def compute_group_outcast_index(args, dict_g_prob, dict_g_size, dict_n_prob, dict_sp, nodes):
	dict_o_index = {}

	for _n in dict_n_prob:
		_o_index = (1+(math.sqrt(dict_g_size[_n]) * dict_g_prob[_n] - dict_n_prob[_n]) /
		(math.sqrt(dict_g_size[_n])    *    dict_g_prob[_n]    +    dict_n_prob[_n])    +
		(args.weigth_factor / nodes * dict_sp[_n])) / (2 + args.weigth_factor)
		dict_o_index[_n] = _o_index

	with open(args.output_f_txt, 'w') as f_o:
		f_o.write("Output group indexes \n")
		f_o.write("\n" + "Nodes: " + str(nodes) + "\n")
		f_o.write("\n" + "Results: \n")
		for _i in sorted(dict_o_index):
			print("Node " + str(_i) + ", Index " + str(dict_o_index[_i]))
		f_o.write("Node " + str(_i) + ", Output index: " + str(dict_o_index[_i]) + "\n")


	with open(args.output_f_nmb, 'w') as f_o:
		for _i in sorted(dict_o_index):
			f_o.write(str(_i) + " " + str(dict_o_index[_i]) + "\n")
```

Ohjelma 5. *Ryhmään kuulumisen indeksin laskentaan käytetty algoritmi*

Algoritmi laskee ryhmään kuulumisen indeksin nopeasti suurillekin verkoille. Rajoittavana tekijänä algoritmissa on muistinhallinta, sillä tämänhetkisessä toteutuksessa talletetaan yksinolemisen todennäköisyyksiä, ryhmäkokoja, ryhmien todennäköisyyksiä sekä leviämistodennäköisyyksiä. Ohjelma 5 esittää toteutusta, jossa $args$ parametriin käyttäjä voi syöttää tässä funktiossa käytetyn painokertoimen (weight_factor) k, joka määrää kumpaa mallia painotetaan enemmän. Parametriin $args$ voidaan syöttää myös tiedostonimet (output_f_txt, output_f_nmb).

5. TULOKSET

Pieniin verkostoihin lukeutuvien testitapausten tarkoituksena on todentaa mallin oikeellisuutta. Malli antaa samankaltaisia tuloksia kuin muut verkostoanalyysin tutkimukset karatekerhon ja hollantilaisopiskelijoiden tapauksissa. Kuvien ja tulosten tarkoituksena on parantaa verkostorakenteiden hahmottamista sekä niiden analysoimista.

Big dataan kuuluvat testitapaukset malli suorittaa. Niissä malli antaa ymmärrettäviä tuloksia. Tarkoituksena niissä oli osoittaa mallin skaalautuvuus suurempiin verkostoihin. Malli skaalautuukin suurempiin verkostoihin kohtalaisesti, mutta niistä saadut tulokset eivät ole aina yhdenmukaisia, jolloin virheen todennäköisyys kasvaa. Merkittävänä tekijänä on vaikutusmallin ajanhetki, jossa verkostoa tarkastellaan. Lyhyillä ajanhetkillä vaikutus on liian vähäistä ja pitkillä vaikutus on ehtinyt levitä liian laajalle. Suuremmissa verkoissa, etenkin sosiaalisen median tapauksessa, käyttäjän ymmärrys verkkoa ja vaikutusmallia kohtaan on tärkeää.

Pelkästään sosiaalisen median avulla tuloksien arviointi on riskialtista, sillä kaikki henkilöt eivät ole rekisteröityneitä sosiaaliseen mediaan. Mallin kannalta henkilöt, jotka eivät ole sosiaalisessa mediassa, ovat näkymättömiä mallille, sillä sosiaalisen verkoston pohjalta kerätty data ei sisällä kyseisiä henkilöitä. Täten mallin ja algoritmin käyttäminen vain sosiaalisen median datan avulla soveltuu ainoastaan sosiaaliseen mediaan, eikä esimerkiksi luokkahuoneen analysoimiseen.

5.1 Pienet verkostot

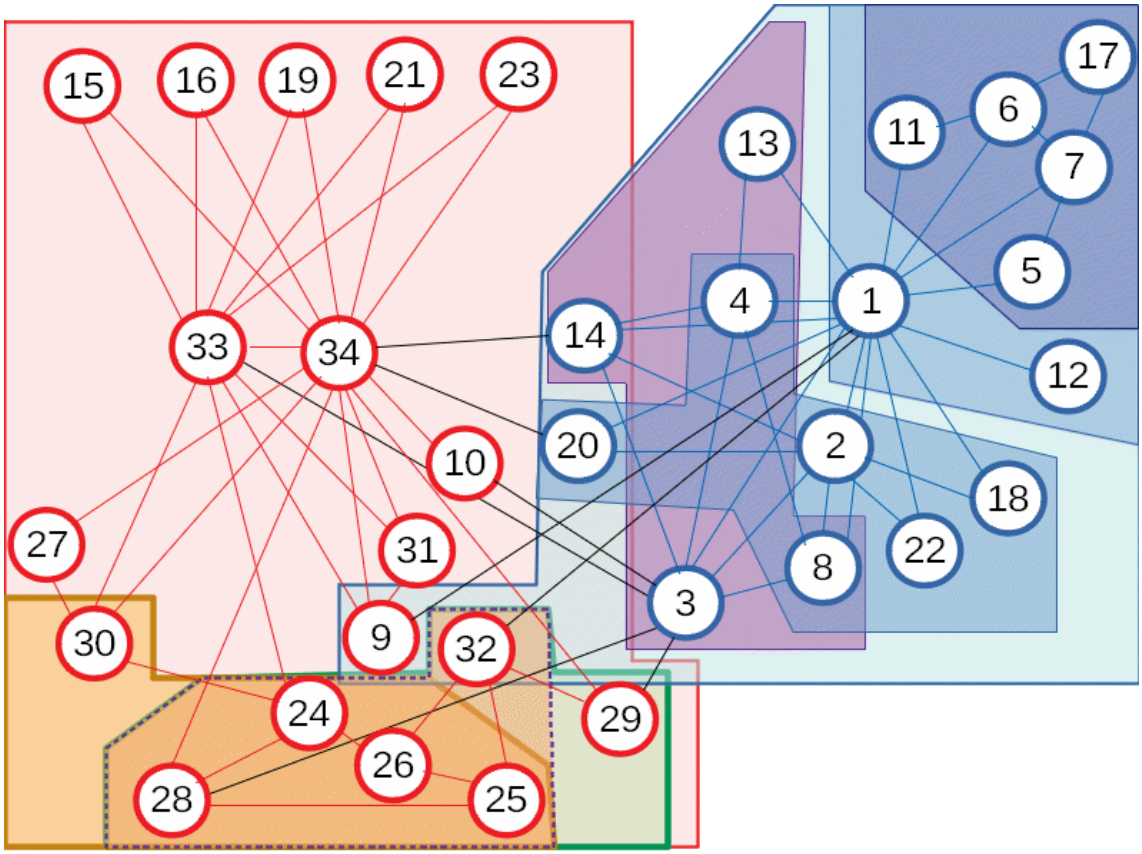
Pienet verkostot sisältävät tässä työssä 30-40 solmua. Niillä testaamisen tarkoituksena on osoittaa ohjelman toimivuus kohderyhmän kokoisilla verkostoilla. Esikoulu, ala- ja yläkoulujen ryhmät sekä erilaiset kerhot ja harrastusryhmät muodostavat lapselle sosiaalisia verkostoja. Mikäli kouluverkostossa lapsella ei ole vaikutusvaltaa ja häntä kiusataan, voi sillä olla vakavia seurauksia tulevaisuuden kannalta.

Ohjelman avulla pienikokoisista verkostoista voi tunnistaa syrjäytyneitä ja mahdollisesti syrjäytyneitä yksilöitä. Tuloksissa niitä tarkastellaan ja visualisoidaan. Testitapaukset on valittu siten, että muutkin tutkijat voivat avoimesti käyttää niitä. Pienten verkostojen hyötynä on niiden nopeat simulaatiot, joita voidaan suorittaa miljoonia kertoja lyhyessä ajassa. Alaluvussa 5.1.1 käsitellään Zacharyn karatekerhoa ja 5.1.2 sisältää hollantilaisopiskelijoista tehdyn verkostoanalyysin.

5.1.1 Zacharyn karatekerho

Zacharyn karatekerho (engl. Zachary's karate club) on yksi yleisimmin käytetyistä esimerkeistä verkkomallinnuksessa. Siinä esiintyy 34 henkilöä, jotka kuuluivat karatekerhoon ja hajaantuivat kahteen ryhmään. Jotta todelliset yhteydet kerhon jäsenten välillä voitiin todeta, tarkasteli sosiologi Wayne Zachary kerhon ulkopuolisia säännöllisiä vuorovaikutuksia jäsenten välillä. Siitä muodostui 78 painottamatonta linkkiä henkilöiden välillä [3].

Tässä tapauksessa sen rakennetta tarkastellaan, jotta syrjäytyneet löydettäisiin ryhmästä. Esimerkki on helppo visualisoida. Verkon rakenne on nähtävissä ja päätelmät ovat helposti analysoitavissa. Lisäksi se tuo luotettavuutta, sillä tämä on todellinen esimerkki. Kuva 9 visualisoi jakoa punaisten ja sinisten solmujen avulla. Algoritmin muodostamia ryhmittymiä esittävät eriväriset alueet. Linkkien värit kuvaavat solmujen ryhmiä. Musta linkki tarkoittaa, että solmut kuuluvat alkuperäisen jaon mukaan eri ryhmiin.



Kuva 9. *Karatekerhon ryhmittymät algoritmin mukaan*

Oletuksena solmu 12 olisi syrjäytynein. Solmu on vain yhteydessä yhteen solmuun, joka on merkittävä. Tämän lisäksi solmut 15, 16, 18, 19, 21, 22 ja 23 olisivat jokseenkin syrjäytyneitä, sillä näillä kaikilla on vain kaksi yhteyttä ja ne osoittavat merkittäviin solmuihin. Merkittävillä solmuilla on paljon muitakin yhteyksiä, joten yksittäisen solmun yhteys ei ole niin voimakas ja täten tiivistä pienryhmää ei muodostu.

Vaikuttamismallin ajamiseen kyseiselle verkolle kului aikaa alle 0,1 sekuntia. Ryhmäjakosimulaatioita tehtiin kymmenessä eri ajossa, joista suurimmassa simulaatioiden määrä oli 10 miljoonaa ja pienimmässä 100 000. Simuloinneissa käytettiin kannettavaa tietokonetta, jossa oli 16 GB keskusmuistia ja prosessorina toimi kaksiytiminen Intel® Core™ i5-6300U 2,4 GHz. Ajoaika kaikissa 14,5 miljoonassa simulaatiossa oli noin 45 minuuttia. Tämä metodi otti huomioon limittäiset ryhmittymät, joka lisäsi ajoaikaa. Syrjäytyneisyyden kaavan soveltaminen laskettuihin tuloksiin vei aikaa alle yhden sekunnin.

Eri simulaatioiden antamat tulokset olivat samoja ryhmään kuulumisen indeksin osalta 1 prosenttiyksikön tarkkuudella. Syrjäytymisindeksin arvo vaihteli maksimissaan 2 prosenttiyksikön välillä. Tämä ei vaikuttanut solmujen riskiin olla syrjäytynyt. Jokainen riskiluokituksen saanut solmu oli tarpeeksi kaukana raja-arvoista $O_i(n)$ 0 % (mahdollisesti syrjäytynyt) ja 50 % (syrjäytynyt).

Taulukko 3. *Karatekerhon minimi- ja maksimitodennäköisyydet mahdolliselle syrjäytymiselle*

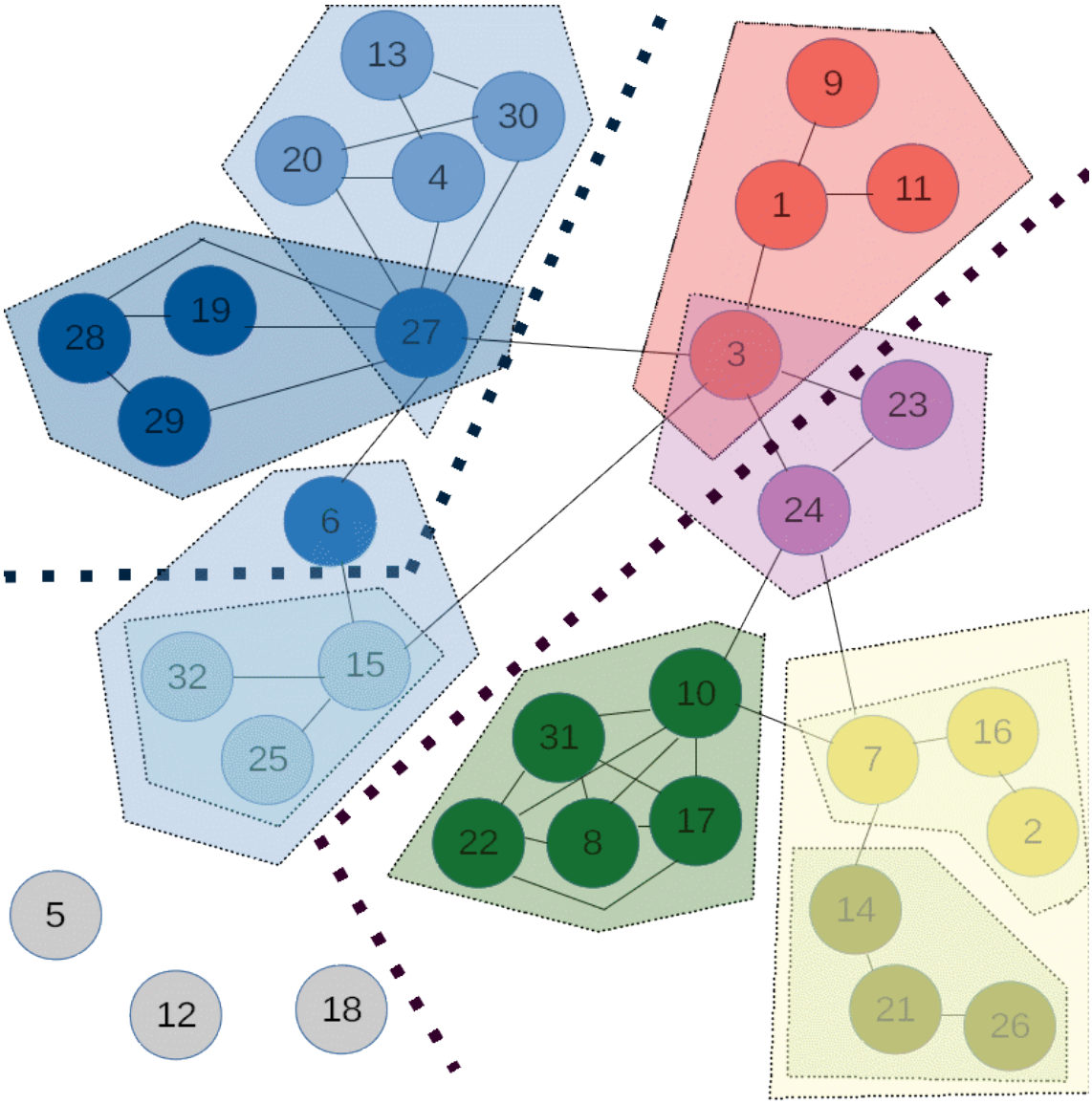
Solmun nro.	Min (%)	Max (%)
10	21.939	22.154
12	44.516	44.755
15	29.736	31.193
16	41.783	41.850
18	17.179	17.537
19	40.039	40.189
20	20.874	21.159
21	32.475	32.999
22	15.247	15.340
23	40.175	40.294
1 - 9, 11, 13, 14, 17, 24 - 34	0.000	0.000

Tuloksissa solmu 12 oli syrjäytynein 46.6 – 46.8 %:n todennäköisyydellä. Tämän jälkeen tulivat solmut 16, 19 ja 23 40.0 – 41.8 %:n todennäköisyyksillä sekä solmut 15 ja 21 29.7 – 33.0 %:n todennäköisyyksillä. Näillä viidellä solmulla on hyvin pitkälti samanlainen sijainti verkostossa, mutta silti todennäköisyydet vaihtelevat paljon. Muita mahdollisesti syrjäytyneitä ovat solmut 10, 18, 20 ja 22 todennäköisyyksillä 15.2 – 22.2 %. Solmu 29 on lähellä syrjäytymisvaaraa, sillä ryhmään kuulumisen indeksi oli välillä 50.1 – 50.4 %, mikä antaa 0 % riskin syrjäytymiselle. Tulokset syrjäytymisvaarassa olevista on nähtävissä taulukossa (Taulukko 3).

5.1.2 Hollantilaiset opiskelijat

Hollantilaiset opiskelijat (engl. Dutch students) on verkkomallinnuksessa käytettävä avoimen datan esimerkki. Tämä on kyselyillä tuotettua tietoa luokkahuoneen rakenteesta opiskelijapiirissä. Materiaalissa esiintyy 32 henkilöä, jotka muodostivat luokan. Heidät

numeroitiin tunnisteilla 1 - 32. Kuva 10 visualisoi verkon rakennetta ja algoritmin havaitsemia ryhmittymiä. Katkoviivat kuvaavat kahtia jaettuja ryhmiä, värilliset taustat pienempiä ryhmittymiä.



Kuva 10. *Hollantilaisopiskelijoiden ryhmittymät algoritmin mukaan*

Oletuksena solmut 5, 12 ja 18 ovat syrjäytyneitä, sillä niillä ei ole yhteyksiä tässä verkostossa. Tämän lisäksi solmut 2, 9, 11, 25, 26 ja 32 olisivat vaarassa syrjäytyä, sillä jokaisella näistä on vain yksi yhteys toiseen solmuun. Näistä solmut 2 ja 26 voivat muodostaa tiiviimmän yhteisön ainoan yhteytensä kanssa, sillä naapurisolmuilla 16 ja 21 on vain yksi yhteys tämän lisäksi.

Vaikuttamismallin ajamiseen kyseiselle verkolle kului aikaa alle 1 sekunti. Simulaatioita ryhmittymiin jakautumisesta tehtiin kymmenessä eri ajossa, jossa suurimmassa simulaatioiden lukumäärä oli 10 miljoonaa ja pienimmässä 100 000. Simulaatioissa käytettiin samaa kannettavaa tietokonetta kuin karatekerhon tapauksessa. Ajoaika kaikissa 14,5

miljoonassa simulaatiossa oli noin 75 minuuttia. Tämä metodi ottaa huomioon limittäiset ryhmittymät, mikä lisää ajoaikaa. Syrjäytyneisyyden kaavan soveltaminen laskettuihin tuloksiin vei aikaa 1,7 sekuntia.

Taulukko 4. *Hollantilaisopiskelijoiden todennäköisyydet mahdolliselle syrjäytymiselle*

Solmun nro.	Min (%)	Max (%)
5	97.917	97.917
9	0.0549	0.1606
11	0.0601	0.2067
12	97.917	97.917
18	97.917	97.917
25	30.811	31.572
32	30.903	31.417
1 - 4, 6 - 8, 10, 13 - 17, 19 - 24, 26 - 31	0.000	0.000

Kuten taulukosta 4 on nähtävissä, selvästi syrjäytyneimpiä ovat solmut 5, 12 ja 18 todennäköisyydellä 97.9 %. Näiden todennäköisyyden poikkeavuus johtuu vaikuttamismallista, jossa vaikuttavuudeksi tulee solmun vaikuttavuus itseensä $1/32$ ja tähän sovelletaan painokerrointa, jossa k:n arvo on 1. Tällöin virheeksi tulee $\frac{1}{32} * \frac{2}{3} = 0.0208 = 2.08\ \%$. Syrjäytymisvaarassa ovat lisäksi solmut 25 ja 32, sillä niiden todennäköisyydet ovat välillä 30.8 – 31.6 %. Mahdollisia syrjäytyviä ovat myös solmut 9 ja 11 selvästi alle 1 %:n todennäköisyydellä.

5.2 Big data

Big datan määritelmät ovat suhteellisia ja vaihtelevat tekijöiden, kuten ajan ja tietotyypin mukaan. Mitä on nyt big dataa, ei tulevaisuudessa välttämättä sitä enää ole, sillä varastointikapasiteetti ja laskentateho kasvavat [29]. Internetin sosiaalisissa verkostoissa yhteisö viittaa käyttäjien alaverkostoon, joka vuorovaikuttaa enemmän toistensa kuin muun verkoston kanssa. Usein miljoonia solmuja ja linkkejä sisältävät verkkoyhteisöt ovat suuria ja ne lasketaan big dataksi.

Big dataa voidaan kuvata seuraavilla ominaisuuksilla: Määrä, valikoima, nopeus, arvo ja todenmukaisuus [21]. Big datan suurimpina hyötyinä ovat virheiden minimointi suuren datamäärän vuoksi, laajempi ymmärrys erilaisista tekijöistä sekä tehokkuuden välttämättömyys mallien sekä ohjelmien luomisessa. Suurimpina ongelmina ovat kompleksiset mallit, joissa laskennallisesti lineaarisuutta $O(N)$ ei saavuteta, datan visualisointi ja hyödyttömän tiedon karsiminen datamassasta.

Big datana pidetään tässä työssä yli 10 gigatavun kokoisia tiedostoja. Tiedostojen alkuperäinen koko voi olla pienempi, mutta vaikutusmallista laskettu data ylittää kyseisen

rajan. Alaluvussa 5.2.1 tarkastellaan Facebook verkkoa, jossa solmuja on 4039. Alaluvussa 5.2.2 Enron sähköpostiverkoston analyysi ja ryhmittymien muodostuminen on avainasemassa. Alalukujen 5.2.1 ja 5.2.2 datasetit ovat avoimesti saatavilla (SNAP) Stanford Network Analysis Project [48].

5.2.1 Facebook

Facebook verkostossa on 4039 solmua ja 88 234 linkkiä. Vaikuttamismallilla sitä on tarkasteltu aiemmin ja koheesio oli kokonaisuudessaan korkealla, kun tarkasteltiin pidemmillä polun pituuksilla [39]. Vaikutusmallin ajaminen vei 50 sekuntia. Ryhmien etsimisessä aikaa kului 49 tuntia yhteensä 260 000 simulaatiolla. Yhdistävän algoritmin ajaminen vei 8 minuuttia. Simulaatiot suoritettiin kolmessa eri ajossa, joissa lukumäärät olivat 10 000, 50 000 ja 200 000 simulaatiota, jonka takia tuloksissa esiintyy pieniä lukumäärällisiä vaihteluja. Vaikutusmallissa tarkasteltiin ajanhetkeä $T = 1.0$, joka kuvaa tilannetta, jossa leviämistä on tapahtunut vaikutusvaltaisempien osalta, mutta kaikkiin solmuihin ei ole vaikutettu.

Taulukko 5. *Facebook verkon todennäköisimmät ryhmäkoot solmuittain*

Ryhmäkoko S_G	**Lukumäärä**
$S_G = 1$	750 - 755
$1 < S_G \leq 5$	1478 - 1483
$5 < S_G \leq 10$	571 - 573
$10 < S_G \leq 100$	1049 - 1069
$100 < S_G \leq 1000$	169 - 183
$S_G > 1000$	0

Taulukko 6. *Facebook verkon solmumääräiset todennäköisyydet syrjäytymiselle*

Syrjäytymisriski $O_i(n)$	**Lukumäärä**
$O_i(n) \geq 50\%$	68
$25\% \leq O_i(n) < 50\%$	54 - 55
$10\% \leq O_i(n) < 25\%$	93 - 96
$0\% < O_i(n) < 10\%$	82 - 86
$O_i(n) = 0\%$	3737 - 3739

Suurimmaksi ryhmäksi, joka oli todennäköisin vaihtoehto solmukohtaisesti, osoittautui 792 solmun kokoinen ryhmä. Toisena suurena ryhmänä esiintyi kooltaan 756 solmua sisältävä ryhmä. Yhteensä 750 – 755 solmua esiintyi yksin todennäköisimpänä vaihtoehtona. Tuloksia ryhmäkoista on esitelty taulukossa (Taulukko 5). Syrjäytyneitä solmuja oli 68 kappaletta ja mahdollisesti syrjäytyneitä solmuja oli 232 - 234 kappaletta, jotka ovat järjestettynä todennäköisyyksien mukaan taulukossa (Taulukko 6).

5.2.2 Enron

Enron on sähköpostiverkko, jonka on alun perin julkaissut Federal Energy Regulatory Commission [48]. Verkko koostuu yksisuuntaisista linkeistä, jotka vastaavat sähköpostin lähettämistä. Verkostossa on 36 692 solmua ja 183 831 linkkiä. Oletuksena sähköpostiverkko on hajanainen. Siellä on mahdollisesti muutamia vaikuttajia, jotka lähettävät ja saavat paljon sähköpostiviestejä. Vaikuttamismallin ajamisessa kului aikaa 24 minuuttia, mikä vastaa lineaarista skaalautuvuutta. Ryhmien etsimisessä ohjelma oli hidas, kuluttaen aikaa 208 tuntia. Ajoja tehtiin yhteensä 10 kappaletta, joissa oli yhteensä vain 3000 simulaatiota, mikä lisää virhemarginaalia paljon. Pienimmässä ajossa oli 100 simulaatiota ja suurimmassa 800 simulaatiota. Algoritmien yhdistämisessä kului aikaa 84 minuuttia. Vaikutusmallissa tarkasteltiin ajanhetkeä $T = 1.0$.

Taulukko 7. *Enron verkon todennäköisimmät ryhmäkoot solmuittain*

Ryhmäkoko S_G	**Lukumäärä**
$S_G = 1$	15 617 – 15 854
$1 < S_G \leq 5$	16 815 – 18 213
$5 < S_G \leq 10$	2 147 – 2 289
$10 < S_G \leq 100$	351 - 690
$100 < S_G \leq 1000$	90 - 819
$S_G > 1000$	36 - 545

Taulukko 8. *Enron verkon solmumääräiset todennäköisyydet syrjäytymiselle*

Syrjäytymisriski $O_i(n)$	**Lukumäärä**
$O_i(n) \geq 50\%$	4 - 5
$25\% \leq O_i(n) < 50\%$	3 039 – 11 789
$10\% < O_i(n) < 25\%$	1 410 – 3 481
$0\% < O_i(n) < 10\%$	605 – 2 916
$O_i(n) = 0\%$	22 883 – 27 483

Solmuja todennäköisimmin yksin esiintyi 15617 – 15854 kappaletta. Suurin ryhmäkoko oli välillä 1291 – 5433 solmua. Tuloksia ryhmäkoista on esitelty taulukossa 7. Syrjäytyneitä solmuja oli yhteensä 4 - 5 kappaletta ja mahdollisesti syrjäytyneitä solmuja oli yhteensä 9 209 – 13 809 kappaletta. Ne ovat järjestettynä todennäköisyysalueiden mukaan taulukossa 8. Paremman analysoinnin kannalta tuloksia pitäisi olla enemmän, mutta ryhmittymien etsimisessä raja hitauden suhteen tulee tämän kokoisessa esimerkissä vastaan. Ajojen tulokset olivat lähellä toisiaan, mikäli simulaatioiden lukumäärä oli sama. Suurimman ja pienimmän ajon tulokset määrittivät raja-arvot lukumäärille. Tuloksia voidaan pitää vain suuntaa antavina, sillä syrjäytyneiden etsiminen on tarkoitus tehdä huomattavasti pienemmälle verkolle. Tässä tapauksessa simulaatioiden lukumäärät olivat niin pienet, että tarkempaa arviota ei pystytä suorittamaan helposti.

6. YHTEENVETO

Kokonaismalli koostuu vaikuttamismallista, yhteisöjen etsimismallista sekä algoritmista, joka yhdistää mallien antamat tulokset todennäköisyyksiksi. Mallit pohjautuvat kirjallisuudessa tutkittuihin menetelmiin ryhmien rakenteista ja keskeisyyden matemaattisista määritelmistä. Vaikuttamismalli sopii hyvin big datan käsittelyyn skaalautuvuutensa vuoksi. Esitetty yhteisöjen etsimiseen käytetty malli soveltuu kohtalaisesti suuriin datamääriin, mutta optimoinnilla sitä voisi soveltaa big dataan. Mallit yhdistävä algoritmi soveltuu hyvin suuriin datamääriin, sillä sen muistinkäyttö on matala ja se suorittaa laskuoperaatioita vähän yksittäistä tulosta kohden.

Yksinkertaisissa verkostoissa, kuten Zacharyn karatekerhossa ja hollantilaisopiskelijoissa tulokset vastaavat oletuksia hyvin pitkälti. Suuremmissa verkostoissa oletuksia ajon tuloksista ei tehty verkoston koon vuoksi. Mallin luotettavuutta on tarkasteltu kriittisesti ja mallissa on todettu puutteita (Taulukko 2) muun muassa vähäisen testiaineiston ja tulosten vertailun osalta. Ajojen antamat tulokset kannustavat lisätutkimukseen ja muiden algoritmien kanssa vertailuun. Kokonaismalli soveltuu verkkojen analysointiin, joissa yhteydet ovat selkeitä. Hyödyllisiä käyttökohteita kokonaismallille olisivat esimerkiksi Facebookin paikkakuntaryhmät, luokkahuoneet ja työpaikat. Suurimman yhteiskunnallisen merkityksen malli saisi, jos sitä käytettäisiin etsimään syrjäytyneitä lapsia luokkahuoneista, jotta syrjäytyminen pystyttäisiin ennaltaehkäisemään jo nuorena.

Tulevaisuuden kehitysideoina on tehostaa yhteisöjen etsimisessä käytettävää mallia tai kehittää sen tilalle toisella lähestymistavalla toteutettava malli. Yhtenä mahdollisuutena olisi soveltaa luotua kokonaismallia todellisuudessa ja analysoida tuloksia eri tieteenalojen ammattilaisten kanssa. Yksittäisiä malleja voi käyttää muihin projekteihin, kuten esimerkiksi rasismin ennaltaehkäisemiseksi, jossa vaikuttamismalli sekä tiiminmuodostamisen malli [45] muutoksineen voisivat antaa mahdolliset ryhmäjaot.

LÄHTEET

[1] J.M. Anthonisse, The rush in a directed graph, Stichting Mathematisch Centrum Mathematisce Besliskunde, Stichting Mathematisch Centrum, 1971, Saatavissa: https://ir.cwi.nl/pub/9791

[2] G.G. Bagnato, J.R.F. Ronqui, G. Travieso, Community detection in networks using self-avoiding random walks, Physica A: Statistical Mechanics and its Applications vol. 505, 2018, pp. 1046–1055, Saatavissa: https://arxiv.org/pdf/1607.08597.pdf

[3] A-L. Barabási, Network science, 2016, Saatavissa: http://networkscience-book.com

[4] B. Barry, Social Exclusion, Social Isolation and the Distribution of Income, CASE Paper 12 London: London School of Economics, 1998, Saatavissa: http://eprints.lse.ac.uk/6516/1/Social_Exclusion,_Social_Isolation_and_the_Distribution_of_Income.pdf

[5] R.F. Baumeister, D.M. Tice, Point-Counterpoints: Anxiety and Social Exclusion, Journal of Social and Clinical Psychology vol. 9 no. 2, 1990, Saatavissa: https://guilfordjournals.com/doi/pdf/10.1521/jscp.1990.9.2.165

[6] A. Bavelas, Communication patterns in task-oriented groups, The Journal of the Acoustical Society of America vol. 22 no. 6, 1950, pp. 725–730

[7] V.D. Blondel, J-L. Guillaime, R. Lambiotte, E. Lefebvre: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment vol. 2008 no. 10, 2008, Saatavissa: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008

[8] L. Bobo, V.L. Hutchings, Perceptions of racial group competition: Extending Blumer's theory of group position to a multiracial social context, American Sociological Review vol. 61 no. 6, 1996, pp. 951–972, Saatavissa: http://www.jstor.org/stable/2096302?seq=1#page_scan_tab_contents

[9] S.P. Borgatti, M.G. Everett, A Graph-theoretic perspective on centrality, Social Networks vol. 28 no. 4, 2006, pp. 466–484, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378873305000833

[10] S.P. Borgatti, M.G. Everett, J.C. Johnson, Analyzing Social Networks (2nd edition), Sage Publications, 2018

[11] U. Brandes, On variants of shortest-path betweenness centrality and their generic computation, Social Networks vol. 30 no. 2, 2008, pp. 136–145, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378873307000731

[12] R. Cadima, J. Ojeda, J.M. Monguet, Social Networks and Performance in Distributed Learning Communities, Journal of Educational Technology & Society vol. 15 no. 4, 2012, pp. 296–304

[13] Y. Chen, P. Zhao, P. Li, Finding Communities by Their Centers, Scientific Reports vol. 6, 2016, Saatavissa: https://www.nature.com/articles/srep24017

[14] A. Chin, Finding Cohesive Subgroups and Relevant Members in the Nokia Friend View Mobile Social Network, International Conference on Computational Science and Engineering vol. 4, 2009. Saatavissa: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5284091

[15] M. A. Christie, J. Glimm, J. W. Grove, D. M. Higdon, D. H. Sharp, M. M. Wood-Schultz, Error Analysis and Simulations of Complex Phenomena, Los Alamos Science no. 29, 2005, Saatavissa: http://www2.stat.duke.edu/~fei/samsi/Readings/DHigdon/lascience.pdf

[16] L.M. Collins, C.W. Dent, Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions, Multivariate Behavioral Research vol. 23 no. 2, 1988, pp. 231–242, Saatavissa: https://doi.org/10.1207/s15327906mbr2302_6

[17] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Physical Review E vol. 72 no. 2, 2005, Saatavissa: https://link.aps.org/doi/10.1103/PhysRevE.72.027104

[18] E. Durkheim, De la division du travail social: étude sur l'orgrganisation des sociétés supérieures, Paris: F. Alcan, 1893, Saatavissa: https://archive.org/details/deladivisiondutr00durkuoft, (Käännetty, L.A. Coser, W.D. Halls, The Division of Labor in Society, The Free Press, 1964, Saatavissa: https://books.google.fi/books?id=B955X3C-9E8C&pg=PR3&hl=fi&source=gbs_selected_pages&cad=2#v=onepage&q&f=false)

[19] EAPN, Social Cohesion at stake: The Social Impact of the Crisis and of the Recovery Package, European Anti-Poverty Network (EAPN) Social Inclusion working group, 2009, Saatavissa: https://www.eapn.eu/wp-content/uploads/Crisis-Report-2009-final.pdf

[20] EAPN, Last Chance for Social Europe?, European Anti-Poverty Network (EAPN) Position Paper on the European Pillar of Social Rights, 2016, Saatavissa: https://www.eapn.eu/wp-content/uploads/2016/09/EAPN-2016-EAPN-Position-European-Pillar-Social-Rights-600.pdf

[21] C.K Emani, N. Cullot, C. Nicolle, Understandable Big Data: A survey, Computer Science Review vol. 17, 2015, pp. 70–81, Saatavissa: https://www.sciencedirect.com/science/article/pii/S1574013715000064

[22] L. Euler, Solutio problematis ad geometriam situs pertinentis, Commentarii academiae scientiarum Petropolitanae 8, 1736, pp. 128–140, Saatavissa: http://euler-archive.maa.org/docs/originals/E053.pdf (Käännetty, N. Biggs, E.K. Lloyd, R.J. Wilson, Graph theory, 1736-1936, Clarendon Press, 1986, pp. 3–11, Saatavissa: https://books.google.fi/books?id=XqYTk0sXmpoC&pg=PA2&hl=fi&source=gbs_selected_pages&cad=2#v=onepage&q&f=false)

[23] T. Falkowski, A. Barth, M. Spiliopoulou, DENGRAPH: A density-based Community Detection Algorithm, International Conference on Web Intelligence (WI'07), 2007, Saatavissa: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4427076

[24] D.R. Forsyth, Group dynamics (7th edition), Cengage Learning, 2018, Saatavissa: https://books.google.fi/books?hl=fi&lr=&id=Ba9EDwAAQBAJ&oi=fnd&pg=PP1&dq&ots=bFvMzulCo6&sig=KyHtMVDHXs5f5EN7Fer5krxBzlg&redir_esc=y#v=onepage&q&f=false

[25] S. Fortunato, Community detection in graphs, Physics Reports vol. 486 no. 3-5, 2010, pp. 75–174, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0370157309002841

[26] S. Fortunato, D. Hric, Community detection in networks: A user guide, Physics Reports vol. 659, 2016, pp. 1–44, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0370157316302964

[27] L. Freeman, A set of measures of centrality based upon betweenness, Sociometry vol. 40 no. 1, 1977, pp. 35–41, Saatavissa: https://www.jstor.org/stable/3033543?seq=1

[28] L. Freeman, Centrality in Social Networks Conceptual Clarification, Social Networks vol. 1, 1978, pp. 215–239, Saatavissa: https://www.bebr.ufl.edu/sites/default/files/Centrality%20in%20Social%20Networks.pdf

[29] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management vol. 35 no. 2, 2015,

pp. 137–144, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0268401214001066

[30] S. Gao, J. Ma, Z. Chen, G. Wang, C. Xing, Ranking the spreading ability of nodes in complex networks based on local structure, Physica A: Statatistical Mechanics and its Applications vol. 403, 2014, pp. 130–147, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378437114001411

[31] J.L. Gross, J. Yellen, Graph theory and its applications (2nd edition), CRC Press, 2005

[32] J.L. Gross, J. Yellen, P. Zhang, Handbook of graph theory (2nd edition), CRC press, 2013

[33] F. Harary, R.Z. Norman, Graph theory as a mathematical model in social science, Research Center for Group Dynamics, University of Michigan, 1953, Saatavissa: http://www.idiosophy.com/wp-content/uploads/2017/07/harary-norman.pdf

[34] C. Haythornthwaite, Social networks and Internet connectivity effects, Information, Communication & Society vol. 8 no. 2, 2005, pp. 125–147

[35] T. Helne, Syrjäytymisen yhteiskunta, Helsingin yliopisto, 2002, Saatavissa: https://helda.helsinki.fi/handle/10138/13240

[36] J. Hills, J. Le Grand, D. Piachaud, Understanding Social Exclusion, Oxford University Press, 2002, Saatavissa: https://books.google.fi/books?hl=fi&lr=&id=pZd9NWN-bEWIC&oi=fnd&pg=PA13&dq=social+exclusion+rates&ots=IjRuD2jWXg&sig=iQ699PN-5sGSTE6SmxCeaP_HZHk&redir_esc=y#v=onepage&q&f=false

[37] J. Holt-Lunstad, T.B. Smith, M. Baker, T. Harris, D. Stephenson, Loneliness and Social Isolation as Risk Factors for Mortality, Persperctives on Psychological Science vol. 10 no. 2, 2015, Saatavissa: http://journals.sagepub.com/doi/pdf/10.1177/1745691614568352

[38] M-H. Hsieh, C.L. Magee, A new method for finding hierarchical subgroups from networks, Social Networks vol. 32 no. 3, 2010, pp. 234–244, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378873310000171

[39] M. Ijäs, J. Levijoki, V. Kuikka, Scalable Algorithm for Computing Influence Spreading Probabilities in Social Networks, Proceedings of the 5th European Conference on Social Media, Limerick Institute of Technology Ireland, 2018, pp.

76–84, Saatavissa: https://books.google.fi/books?hl=en&lr=&id=b09mDwAAQBAJ

[40] S. Ispa-Landa, Gender, Race, and Justifications for Group Exclusion: Urban Black Students Bussed to Affluent Suburban Schools, Sociology of Education vol. 86 no. 3, 2013, Saatavissa: http://journals.sagepub.com/doi/full/10.1177/0038040712472912

[41] S. Keltanen, Sosiaalisen verkon visuaalisen analysoinnin selkiyttäminen vähentämällä linkkien ja solmujen määrää, Jyväskylän yliopisto, 2016, Saatavissa: https://jyx.jyu.fi/dspace/bitstream/handle/123456789/48538/URN%3aNBN%3afi%3ajyu-201601311353.pdf?sequence=1

[42] D. Khullar, How social isolation is killing us, The New York Times (Dec. 22. 2016), 2016, Saatavissa: https://www.nytimes.com/2016/12/22/upshot/how-social-isolation-is-killing-us.html

[43] J. Korhonen, Reitinhaku pienissä maailmoissa, Helsingin yliopisto, 2009, Saatavissa: https://www.cs.helsinki.fi/u/htoivone/teaching/seminaariK09/janne_korhonen.pdf

[44] V. Kuikka, Influence Spreading Model Used to Community Detection in Social Networks, International Conference on Complex Networks and their Applications Complex Networks 2017: Complex Networks & Their Applications VI, Springer, 2017, pp. 202–215

[45] V. Kuikka, J. Latikka, Mathematical method for Selecting Team Members from a Social Network, International Journal of Engineering Research and Development vol. 14 no. 2, 2018, pp. 1–15, Saatavissa: https://www.researchgate.net/profile/Vesa_Kuikka/publication/323445107_Mathematical_Method_for_Selecting_Team_Members_from_a_Social_Network/links/5a96515d45851535bcdcca8c/Mathematical-Method-for-Selecting-Team-Members-from-a-Social-Network.pdf

[46] V. Kuikka, Terrorist Network Analyzed with an Influence Spreading Model, International Workshop on Complex Networks CompleNet 2018: Complex Networks IX, Springer, 2018, pp. 185–197

[47] R. Lambiotte, V. Salnikov, M. Rosvall, Effect of memory on the dynamics of random walks on networks, Journal of Complex Networks vol. 3 no. 2, 2015, pp. 177–188, Saatavissa: https://academic.oup.com/comnet/article/3/2/177/375705

[48] J. Leskovec, A. Krevl, SNAP Datasets: Stanford Large Network Dataset Collection, 2014, Saatavissa: http://snap.stanford.edu/data

[49] J. Letkowski, Applications of the Poisson probability distribution, Proceedings of Academic and Business Research Institute Conference, 2012, pp. 1–11, Saatavissa: https://www.researchgate.net/profile/Jerzy_Letkowski/publication/272170277_Developing_Poisson_probability_distribution_applications_in_a_cloud/links/54dd40ba0cf282895a3b4e6e/Developing-Poisson-probability-distribution-applications-in-a-cloud.pdf

[50] K. Lewin, Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change, Human Relations vol. 1 no. 1, 1947, pp. 5–41, Saatavissa: http://journals.sagepub.com/doi/pdf/10.1177/001872674700100103

[51] J. Liu, Q. Xiong, W. Shi, X. Shi, K. Wang, Evaluating the importance of nodes in complex networks, Physica A: Statistical Mechanics and its Applications vol. 452, 2016, pp. 209–219, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378437116002156

[52] O. Mason, M. Verwoerd, Graph theory and networks in Biology, IET Systems Biology vol. 1 no. 2, 2007, pp. 89–119

[53] J.M. McPherson, P.A. Popielarz, S. Drobnic, Social Networks and organizational dynamics, American Sociological Review vol. 57 no. 2, 1992, pp. 153–170, Saatavissa: https://www.jstor.org/stable/2096202?seq=1

[54] J. Moody, D. McFarland, S. Bender-deMoll, Dynamic Network Visualization, American Journal of Sociology vol. 110 no. 4, 2005, pp. 1206–1241

[55] J. Moody, Diffusion and Visualization in Dynamic Networks, Sunbelt XXVI Social Network Conference, (Vancouver Canada), 2006, Saatavissa: https://slideplayer.com/slide/8589712/

[56] J.L. Moreno, H.H. Jennings, Statistics of Social Configurations, Sociometry vol. 1 no. ¾, 1938, pp. 342–374, Saatavissa: http://www.jstor.org/stable/2785588?seq=1

[57] P. Myrskylä, Hukassa – Keitä ovat syrjäytyneet nuoret, Elinkeinoelämän Valtuuskunta (EVA) analyysi no. 19, 2012, Saatavissa: https://www.eva.fi/wp-content/uploads/2012/02/Syrjaytyminen.pdf

[58] M.E.J. Newman, The structure and function of complex networks, Society for Industrial Applied Mathematics (SIAM) review vol. 45 no. 2, 2003, pp. 167–256, Saatavissa:

https://epubs.siam.org/doi/pdf/10.1137/S003614450342480?xid=PS_smithson-ian&

[59] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E vol. 69 no. 2, 2004, Saatavissa: https://link.aps.org/doi/10.1103/PhysRevE.69.026113

[60] M.E.J. Newman, A measure of betweenness centrality based on random walks, Social Networks vol. 27 no. 1, 2005, pp. 39–54, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378873304000681

[61] M.E.J. Newman, T.P. Peixoto, Generalized Communities in Networks, Physical Review Letters vol. 115 no. 8, 2015, Saatavissa: https://link.aps.org/doi/10.1103/PhysRevLett.115.088701

[62] M.E.J. Newman, G. Reinert, Estimating the Number of Communities in a Network, Physical Review Letters vol. 117 no. 7, 2016, Saatavissa: https://link.aps.org/doi/10.1103/PhysRevLett.117.078301

[63] W.L. Oberkampf, S.M. DeLand, B.M. Rutherford, K.V. Diegert, K.F. Alvin, Error and uncertainty in modeling and simulation, Reliability Engineering & System Safety vol. 75 no. 3, 2002, pp. 333–357, Saatavissa: https://www.sciencedirect.com/science/article/pii/S095183200100120X

[64] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community Detection in Social Media, Data Mining and Knowledge Discovery vol. 24 no. 3, 2012, pp. 515–554, Saatavissa: https://link.springer.com/article/10.1007/s10618-011-0224-z

[65] The Pew Forum, Global Restrictions on religion, Pew Research Center, 2009, Saatavissa: http://www.pewforum.org/files/2009/12/restrictions-fullreport.pdf

[66] M. Pulli, Virtaustekniikka, Tampere: Tammertekniikka, 2009, s. 145

[67] A.R. Radcliffe-Brown, On Social Structure, The Journal of the Royal Anthropological Institute of Great Britain and Ireland vol. 70 no. 1, 1940, pp. 1–12, Saatavissa: https://www.jstor.org/stable/2844197?seq=1

[68] M. Roswall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proceedings of National Academy of Sciences of the United States of America (PNAS) vol. 105 no. 4, 2008, pp. 1118–1123, Saatavissa: https://doi.org/10.1073/pnas.0706851105

[69] G. Sabidussi, The centrality index of a graph, Psychometrika vol. 31 no. 4, Springer, 1966, pp. 581–603, Saatavissa: https://doi.org/10.1007/BF02289527

[70] J. Scott, Social Network Analysis: A Handbook (2nd edition), Sage publications, 2000

[71] J. Scott, P.J. Carrington, The SAGE Handbook of Social Network Analysis, Sage publications, 2011

[72] J. Scott, Social Network Analysis (4th edition), Sage Publications, 2017, Saatavissa: https://books.google.fi/books?hl=fi&lr=&id=i5EmDgAAQBAJ&oi=fnd&pg=PP1&dq=social+networks+analysis&ots=DCQ9hSSdf4&sig=PGEL2q0pQ9qZ_sYP0wOiKuSN1HM&redir_esc=y#v=onepage&q=social%20networks%20analysis&f=false

[73] Social Exclusion Unit, Preventing Social Exclusion, 2001, Saatavissa: http://www.bristol.ac.uk/poverty/downloads/keyofficialdocuments/Preventing%20Social%20Exclusion.pdf

[74] W.G. Stephan, C.W. Stephan, Intergroup relations, Social Psychology Series, 2018, Saatavissa: https://content.taylorfrancis.com/books/download?dac=C2017-0-71667-6&isbn=9780429968280&format=googlePreviewPdf

[75] H. Sun, E. Ch'ng, X. Yong, J. M. Garibaldi, S. See, D. Chen, A fast community detection method in bipartite networks by distance dynamics, Physica A: Statistical Mechanics and its Applications vol. 496, 2018, pp. 108–110, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0378437117313481

[76] L.Y. Tang, S.N. Li, J.H. Lin, Q. Guo, J.G. Liu, Community structure detection based on the neighbor node degree information, International Journal of Modern Physics C vol. 27 no. 4, 2016

[77] P. Tsakloglou, F. Papadopoulos, Aggregate level and determining factors of social exclusion in twelve European countries, Journal of European Social Policy vol. 12 no. 3, 2002, pp. 211–225, Saatavissa: http://journals.sagepub.com/doi/pdf/10.1177/0952872002012003394

[78] G.K.F. Tso, K.K.W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy vol. 32 no. 9, 2007, pp. 1761–1768, Saatavissa: https://www.sciencedirect.com/science/article/pii/S0360544206003288

[79] F. Tönnies, Gemeinschaft und Gesellschaft, Leipzig: Fues's Verlag, 1887, (Käännetty, C.P. Loomis, Community and Society, East Lansing: Michigan State University Press, 1957)

[80] C. Westphal, Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies. CRC Press, 2008

[81] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, ACM Computing Surveys (CSUR) vol. 45 no. 4, 2013, Saatavissa: https://dl.acm.org/citation.cfm?id=2501657

[82] D. Yashenkova, Violation of transgender people's rights in Russia, Transgender Legal Defense Project, 2016, Saatavissa: http://pravo-trans.eu/files/violation_of_the_rights_of_transgender_people_in_Russia-en.pdf

LIITE A: LEVIÄMISTODENNÄKÖISYYKSIEN LASKENTA KAIKKIALTA YHTEEN SOLMUUN (C++ TOTEUTUS)

```
void solve_wide(const Network &net, const Config &conf,
                result_callback print_res)
{
    int L = conf.max_length;
    int N = net.nodeCount();
    Spinner spinner(std::clog);
    spinner.set_limit((L+1)*conf.start_nodes.size());
    auto nodes = net.getNodes();
    auto stride = static_cast<int>(conf.times_stride);
    aligned_vector<double> buf[2];
    aligned_vector<double> model_values(stride, 0.0);
    aligned_vector<double> inv_model_values(stride, 0.0);
    spinner.spin(0);
    for(auto target : conf.start_nodes) {
        buf[0].assign(N*stride, 0.0);
        buf[1].assign(N*stride, 0.0);
        for(auto depth : boost::irange(L, -1, -1)) {
            int curr = depth % 2;
            int prev = (curr+1)%2;
            for(auto t : boost::irange(0, stride)) {
                model_values[t] = poisson(depth, conf.times[t], conf.lambda);
                inv_model_values[t] = 1 / model_values[t];
            }
            buf[curr].assign(N*stride, 0.0);
            std::copy_n(begin(model_values), stride,
                        begin(buf[curr])+target*stride);
            for(auto from_index : boost::irange(0, N)) {
                for(auto link : nodes[from_index].neighbours) {
                    int to_index = link.targetIndex;
                    accumulate_to_next(&buf[prev][to_index*stride],
                             &buf[curr][from_index*stride],
                             link.weight*nodes[to_index].weight,
                             inv_model_values);
                }
            }
            spinner.spin();
        }
        print_res(buf[0], target);
    }
    std::cout.flush();
}
```